MIS772

Predictive Analytics

Welcome and Introduction

Textbook by Vijay Kotu and Bala Deshpande,
Data Science: Concepts and Practice, 2nd ed, Elsevier, 2019.

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

Deakin University CRICOS Provider Code: 00113B

Welcome and Intro

- Welcome
- About our team
- Predictive analytics
- Our tool
- Analytics process
- Example with a data set
- Exploring and visualising data
- Hands-on demonstration

DEAKIN BUSINESS SCHOOL

About our Teaching Team

Arman Kaldi
(e-HRM)
a.kaldi@deakin.edu.au

Plus a strong team of expert tutors

Dr Kaushi Nallaperuma
(Marketing Analytics)

Dr Joerin Motavallian
(Supply Chain Analytics)

Dr Mina Roshan Kokabha
(Business Analytics)

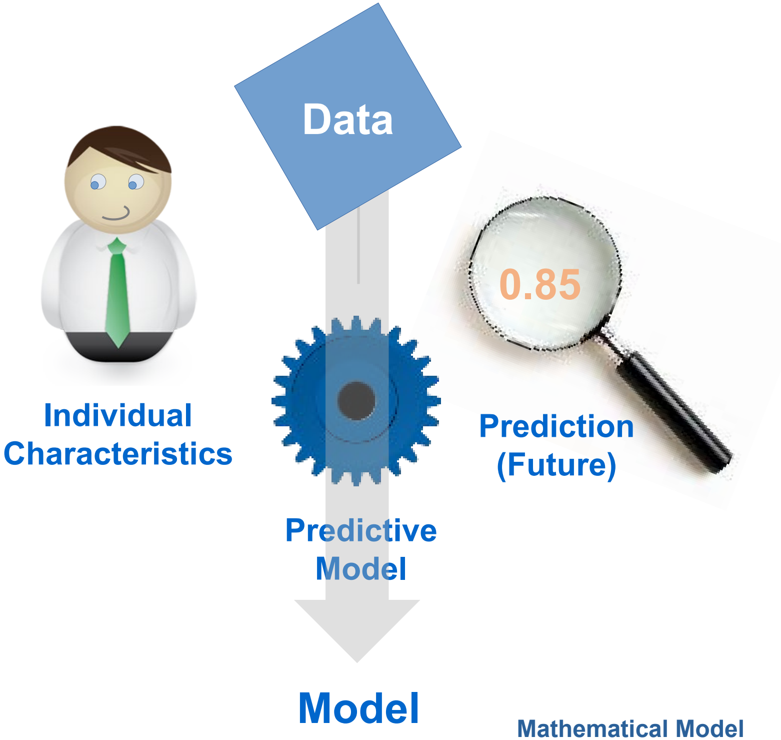
Let’s get started …

Our unit site, learning outcomes, schedule of topics, assessments, learning resources, support

Mentimeter

Predictive Analytics
Data Mining + Model Building

- Predictive Analytics is the practice of extracting (or mining) information from existing data sets to determine patterns and predict future outcomes and trends. (Webopedia)
- It focuses on building models from data – mathematical, machine learning or visual.
- Data model is an abstract description of data, representing its most important characteristics and their patterns.
- A model allows us to gain insights about data from the past, present and future.
- Commonly, the same model can be used for explanation, decision support and prediction.
- Data sets used in model building are often very large, sometimes they must be very large for the model to be of a high quality.
- Data may be collected by an organisation, purchased from others or obtained from open data repositories.
- Special techniques need to be used when data is very small.



Mathematical Model
Machine Learning Model
Visual Model

Also see KD 1.2.2 on Data Modelling

DEAKIN BUSINESS SCHOOL

Tools & Methods

Open Source Tools:

- R with R Studio
- Python / Anaconda with Spyder
- WEKA

O/S Machine Learning Tools:

- Tensorflow, Sagemaker, PyTorch, MXNet, H2O.ai, ChatGPT

Commercial / Community Tools:

- RapidMiner Studio
- KNIME Analytics Platform

Commercial Tools:

- SAS Viya
- IBM SPSS Modeler
- Microsoft Azure ML
- MathWorks Matlab

Approaches to Analytic Problem Solving:

- Selection of statistical methods
  - Linear regression
  - Logistic regression
  - Time series analysis
  - Anomaly detection
  - General linear models
  - Bayesian modelling
  - Association analysis
- Selection of machine learning methods
  - Lazy methods (k-NN)
  - Decision trees and forests
  - Cluster analysis
  - Support vector machines
  - Neural networks and deep learning
  - Genetic algorithms
- Text Processing
  - Text mining
  - Topic and cluster analysis
  - Sentiment analysis
  - Natural Language Processing

Blue – taught in this unit.
Black – taught elsewhere.

Approaches not covered in this unit are taught elsewhere, e.g.

- Marketing Analytics
- Advanced AI for Business
- Modern Data Science

Also see KD 1.1 to
compare AI, ML and DS
Then check KD 1.4-1.5 on
Tasks and Algorithms

DEAKIN
BUSINESS
SCHOOL

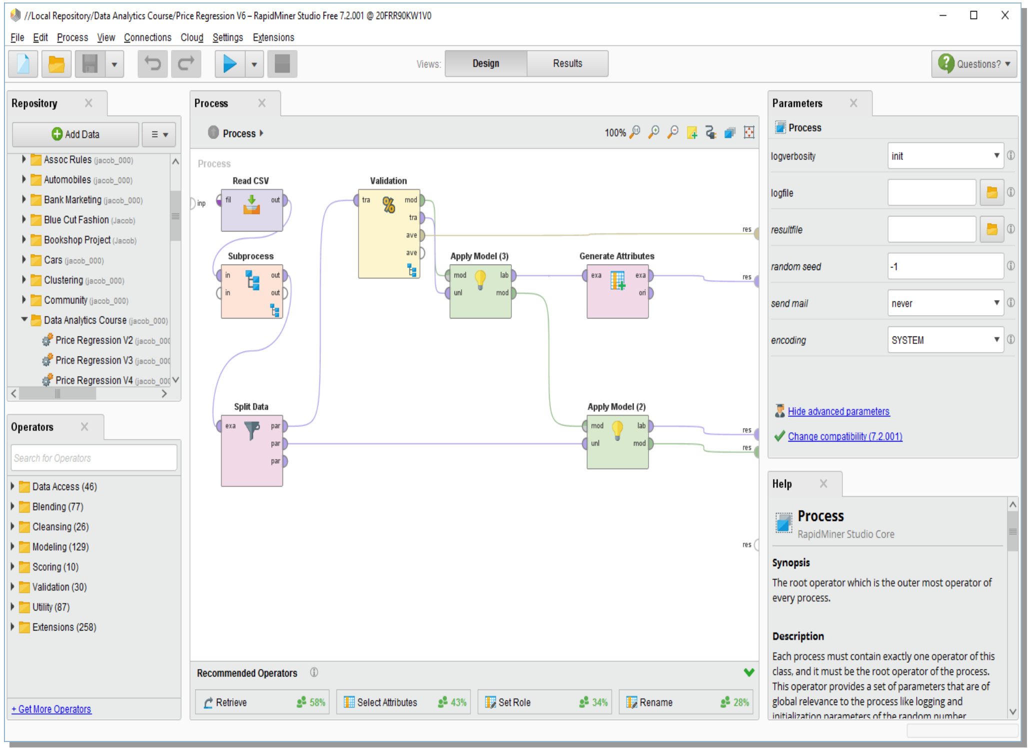
RapidMiner Studio

Altair AI Studio 2024 (RapidMiner) – Your Tool

- **Install RM Studio**

- Do not use “free” (or “trial”) versions of RM as their functionality is limited, e.g. max 1,000 records!
- You also need extensions

More info in the first lab!



logverbosity: init

logfile:

resultfile:

random seed: -1

send mail: never

encoding: SYSTEM

Process

RapidMiner Studio Core

Synopsis

The root operator which is the outer most operator of every process.

Description

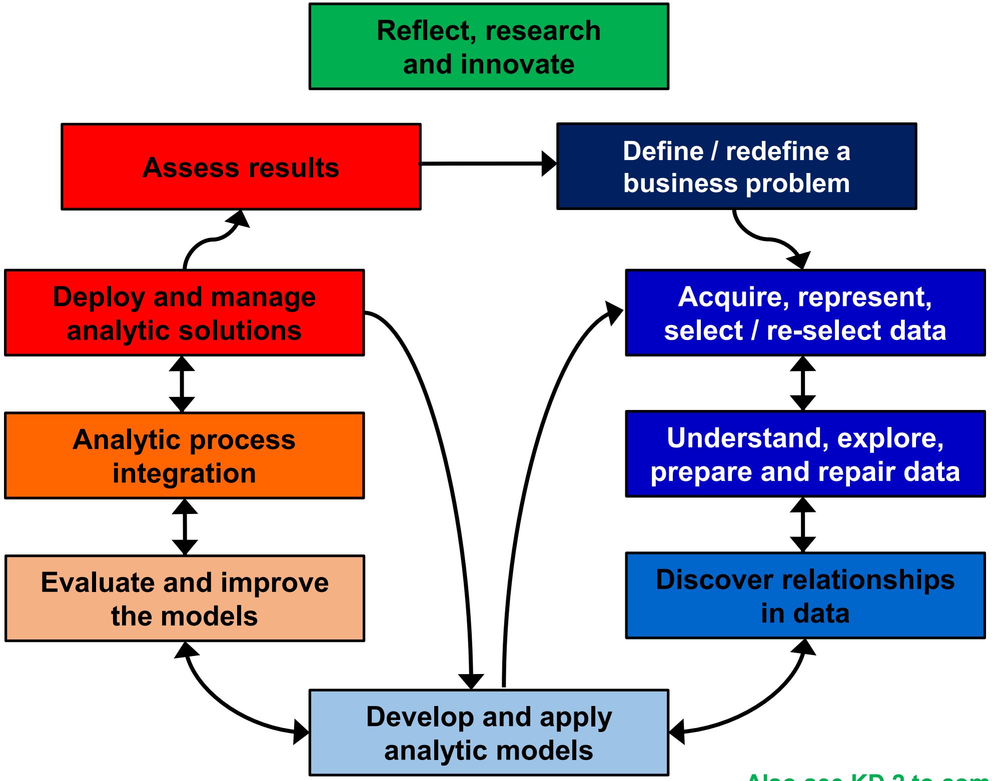
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization parameters of the random number

Also see KD 15 on RapidMiner Studio, its user interface, functionality and use

DEAKIN BUSINESS SCHOOL

Analytics Process

Abstract



Reflect, research and innovate
Assess results
Define / redefine a business problem
Deploy and manage analytic solutions
Acquire, represent, select / re-select data
Analytic process integration
Understand, explore, prepare and repair data
Evaluate and improve the models
Discover relationships in data
Develop and apply analytic models

Also see KD 2 to compare with CRISP-ML and SEMMA process defs

Define a business problem in analytic terms – Formulate a business problem and specify requirements for its solution in terms of insights to be generated, as well as, decisions that need to be made and turned into business actions.

Understand and prepare data – Select a data sample; explore and understand attributes characteristics; deal with missing values and outliers; clean, transform, convert and select attributes to suit the modelling approach.

Discover relationships in data – Explore, visualise and understand relationships between attributes; determine labels and their predictors.

Develop and apply analytic models – Build a collection of predictive and/or explanatory analytic models using statistical, data mining or text mining algorithms. Study the models' characteristics.

Evaluate and improve the models – Validate and test each model for its ability to predict or explain attribute values; evaluate each model performance; tune the model to optimise its performance; compare the models; select the optimal; interpret and report all results.

Analytic process integration – Integrate pre-processing, exploratory and predictive analytic elements and visualisations into a complete analytic process.

Deploy, manage and assess analytic solutions – Embed the final analytic process in a business application; apply the process to live data; use the results to support business decisions and actions; measure and assess the model performance on real data; reflect, research and innovate.

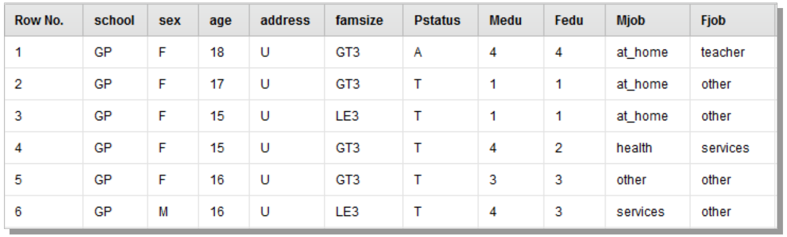
Analytics Process

In Practice

Analytic process describes a reusable *workflow* of data and control through analytic operators. Such operators are responsible for data preparation, creation of a data model, its validation or reporting of results. Some analytics tools focus on workflow management, rather than simply model development, e.g. RapidMiner, Azure ML Studio, SPSS Modeler or SAS Viya.

Define a problem

Observe, acquire and represent data



Row No.	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
1	GP	F	18	U	GT3	A	4	4	at_home	teacher
2	GP	F	17	U	GT3	T	1	1	at_home	other
3	GP	F	15	U	LE3	T	1	1	at_home	other
4	GP	F	15	U	GT3	T	4	2	health	services
5	GP	F	16	U	GT3	T	3	3	other	other
6	GP	M	16	U	LE3	T	4	3	services	other

Prepare, select, clean and transform attributes

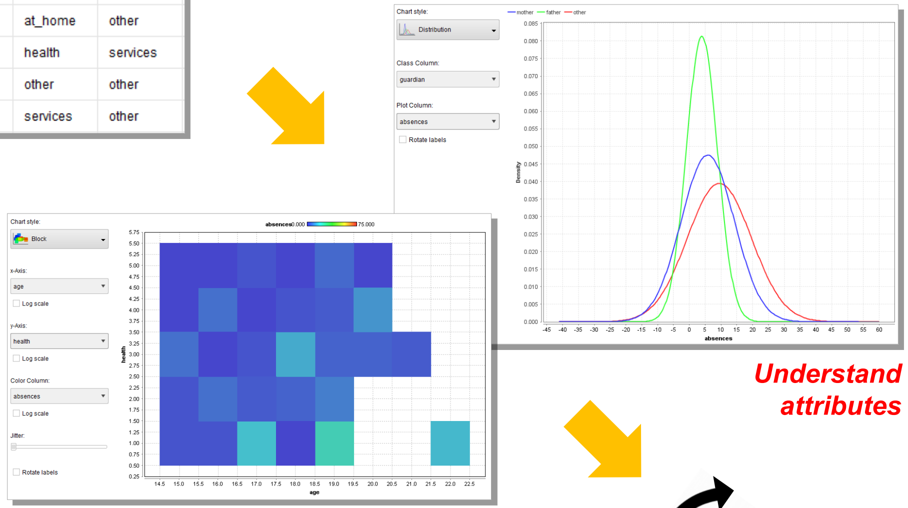


Chart style: Distribution

Class Column: guardian

Plot Column: absences

□ Rotate labels

Chart style: Block

x-Axis: age

□ Log scale

y-Axis: health

□ Log scale

Color Column: absences

□ Log scale

Jitter:

□ Rotate labels

Understand attributes

Discover relationships

Build, evaluate and deploy models

Case and Data Set

Source:
http://archive.ics.uci.edu/ml/datasets/Student+Performance

- **The data used in this presentation includes student achievement in secondary Maths education of two Portuguese schools.**
- **The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires.**
- **The task is to predict student's success in Maths and provide assistance early.**
- **Students' results G1 and G2 correspond to the 1st and 2nd period grades. G3 is the final year grade issued at the 3rd period. G3 has a strong correlation with attributes G1 and G2. It is more difficult to predict G3 without G1 and G2, but such prediction is much more useful.**

Data set attribute Information (";" separated):

01 school - student's school (GP/MS)
02 sex - student's sex (F/M)
03 age - student's age (15..22)
04 address - urban or rural address (U/E)
05 famsize - family size ≤3 or >3 (LE3/GT3)
06 Pstatus - parent's living together or apart (T/A)
07 Medu - mother's education (0..4)
08 Fedu - father's education (0..4)
09 Mjob - mother's job (label)
10 Fjob - father's job (label)
11 reason - reason to choose this school (label)
12 guardian - student's guardian (label)
13 traveltime - home to school travel time (1..4)
14 studytime - weekly study time (1..4)
15 failures - number of past class failures (1..3, or 4)
16 schoolsup - extra educational support (yes/no)
17 famsup - family educational support (yes/no)
18 paid - extra paid classes (yes/no)
19 activities - extra-curricular activities (yes/no)
20 nursery - attended nursery school (yes/no)
21 higher - wants to take higher education (yes/no)
22 internet - Internet access at home (yes/no)
23 romantic - with a romantic relationship (yes/no)
24 famrel - quality of family relationships (1..5)
25 freetime - free time after school (1..5)
26 goout - going out with friends (1..5)
27 Dalc - workday alcohol consumption (1..5)
28 Walc - weekend alcohol consumption (1..5)
29 health - current health status (1..5)
30 absences - number of school absences (0..99)
31 G1 - first period grade (0..20)
32 G2 - second period grade (0..20)
33 G3 - final grade (0..20)

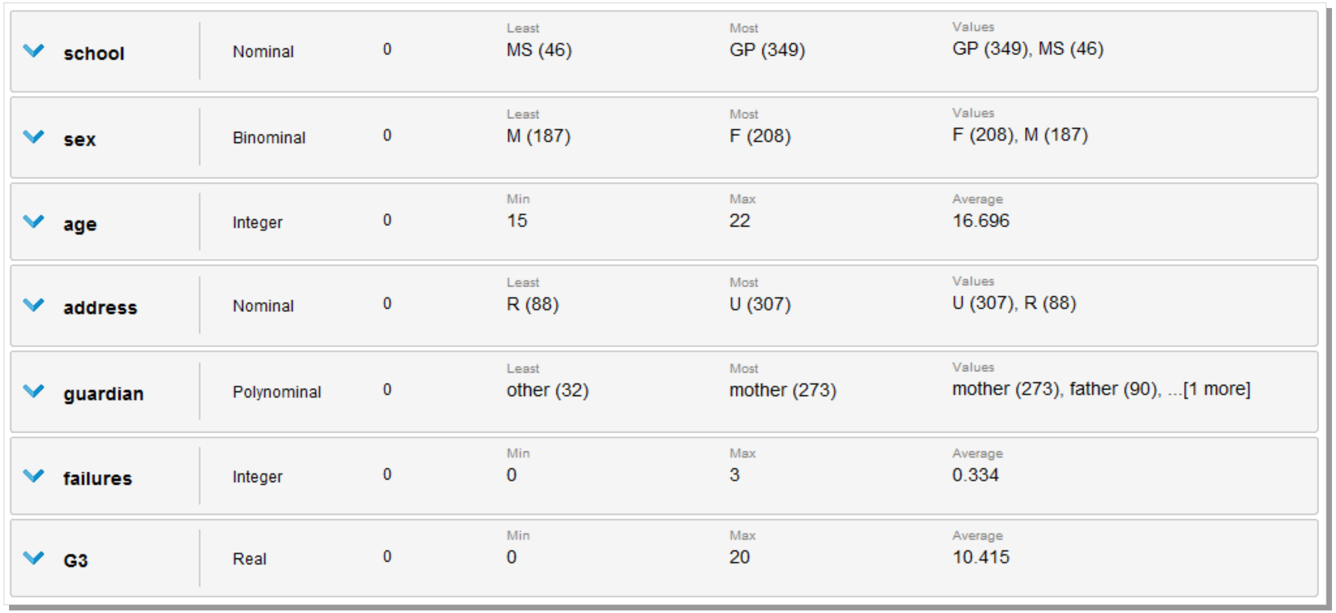
Also see KD 3.2-3.3 on data types and stats

Variables & Attributes

Variables represent the values of the observed *attributes*, i.e. characteristics of people, objects and events. Many analytics systems refer to them as variables, others like RapidMiner, refer to them as attributes. There are many types of attributes:

- *Categorical / Nominal* which describe qualities, e.g.
  - *Binomial* to allow only two possible values, e.g. in this data set "M" and "F" a student's sex
  - *Polynomial* to allow more than two possible values, e.g. "mother", "father" or "other" of the high school student's guardian

- *Numerical* which describe quantities or amounts, e.g.
  - *Discrete / Ordinal* to allow counting measures, e.g. the number of students or the age in years
  - *Continuous* to allow use of real numbers, e.g. the assignment or exam mark
- *Special*, e.g. Date, Time or Text



school	Nominal	0	Least MS (46)	Most GP (349)	Values GP (349), MS (46)
sex	Binominal	0	Least M (187)	Most F (208)	Values F (208), M (187)
age	Integer	0	Min 15	Max 22	Average 16.696
address	Nominal	0	Least R (88)	Most U (307)	Values U (307), R (88)
guardian	Polynominal	0	Least other (32)	Most mother (273)	Values mother (273), father (90), ...[1 more]
failures	Integer	0	Min 0	Max 3	Average 0.334
G3	Real	0	Min 0	Max 20	Average 10.415

Assumptions

In deciding on the type of an attribute, it is important to understand business and legal assumptions as well as user requirements rather than political or cultural biases of the analyst. And so in this Portuguese data, as collected at the time (before EU was founded), sex was defined as having only two possible values. This may change as the laws and culture are revised.

Data Exploration & Visualisation

- Data visualisation and analytics can be used to investigate properties of individual attributes.
- More importantly they can also be used to explore various kinds of relationships between attributes.
- Relationships can be observed by viewing distribution or aggregation of pairs of values, e.g. using a stacked bar graph or scatter plots (and more).

Data visualisation serves as an extension of the analyst's memory and assist forming intuition about data, models and their performance.

Also see KD 3.4 on data visualisation, Check out all KD 3

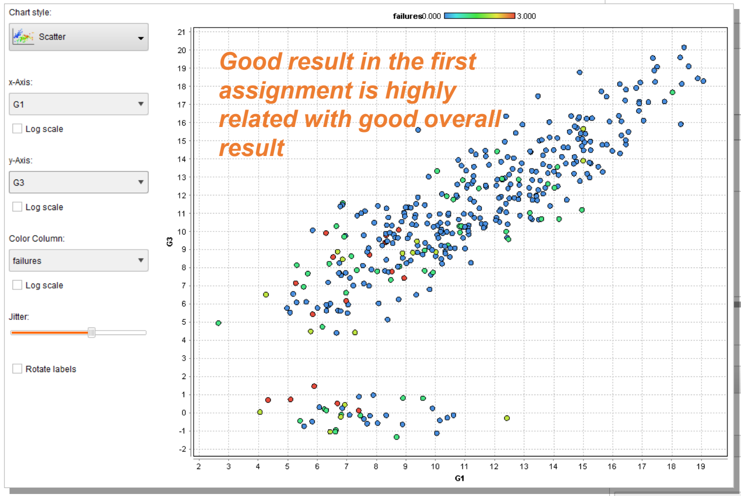


Chart style: Scatter

x-Axis: G1

□ Log scale

y-Axis: G3

□ Log scale

Color Column: failures

□ Log scale

Jitter:

□ Rotate labels

Good result in the first assignment is highly related with good overall result

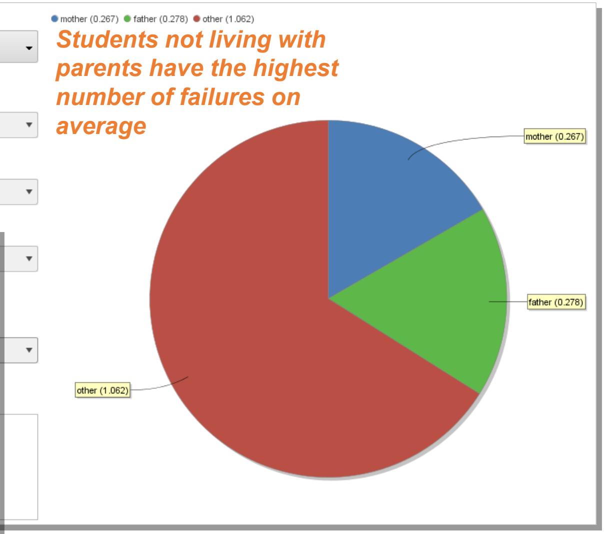
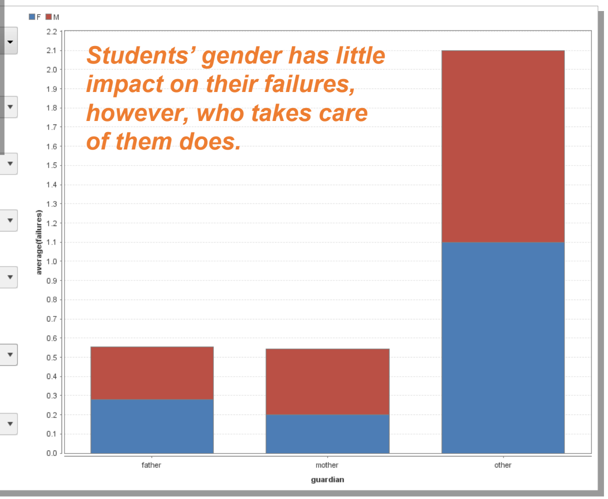


Chart style: Pie

Group-By Column: guardian

Legend Column: None

Students not living with parents have the highest number of failures on average


sex

Legend Column: None

Value Column: average(failures)

□ Absolute values

Aggregation: none

□ Use Only Distinct

□ Rotate labels

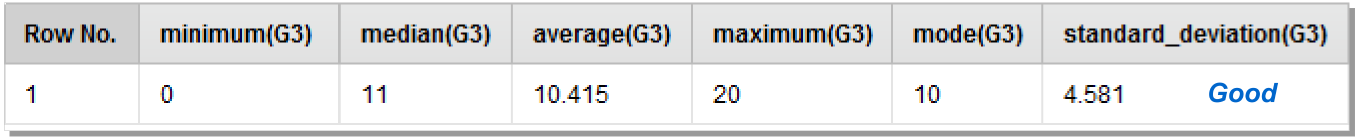
vertical

Students' gender has little impact on their failures, however, who takes care of them does.

More complex visualisations and analytic techniques may assist exploration of multi-dimensional relationships, e.g. PCA with scatter plots.

Data Visualisation

Not always visualisation is a better way of data exploration. Statistics are usually more precise. However, select the level of data aggregation and statistical analysis that is most suitable for your problem. Basic statistics provide single value aggregations of attribute properties, box plots and histograms provide more information about data. Experiment with both visualisation and statistics to gain best insights!



Row No.	minimum(G3)	median(G3)	average(G3)	maximum(G3)	mode(G3)	standard_deviation(G3)
1	0	11	10.415	20	10	4.581

Good

Equivalent variable statistics

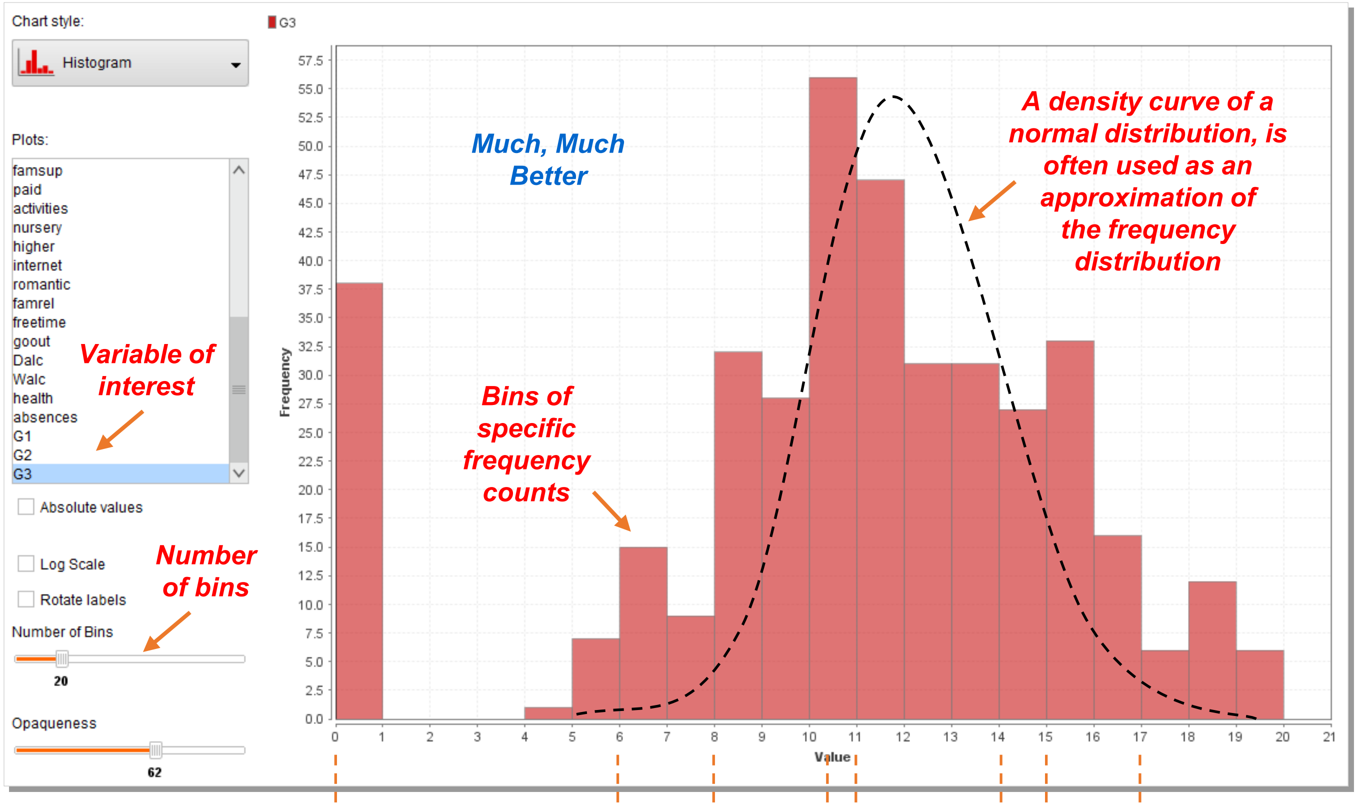


Chart style: Histogram

Plots: G3

☐ Absolute values

☐ Log Scale

☐ Rotate labels

Number of Bins: 20

Opaqueness: 62

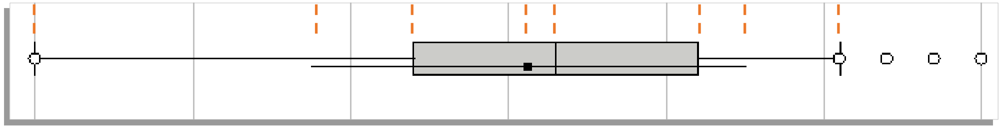
Much, Much Better

Variable of interest

Number of bins

Bins of specific frequency counts

A density curve of a normal distribution, is often used as an approximation of the frequency distribution



Equivalent variable box plot

Better

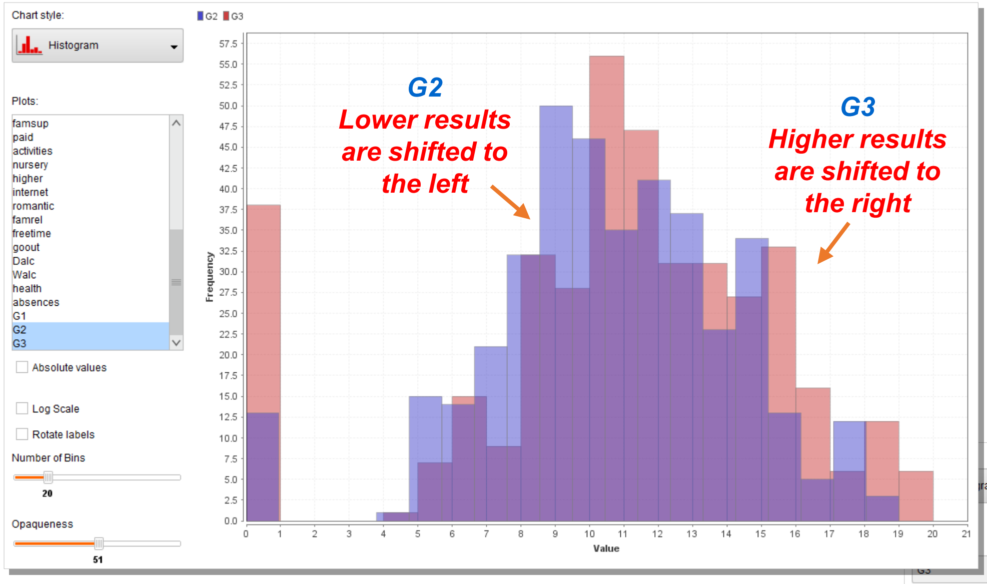


Chart style: Histogram

Plots:
famsup
paid
activities
nursery
higher
internet
romantic
famrel
freetime
goout
Dalc
Walc
health
absences
G1
G2
G3

☐ Absolute values

☐ Log Scale

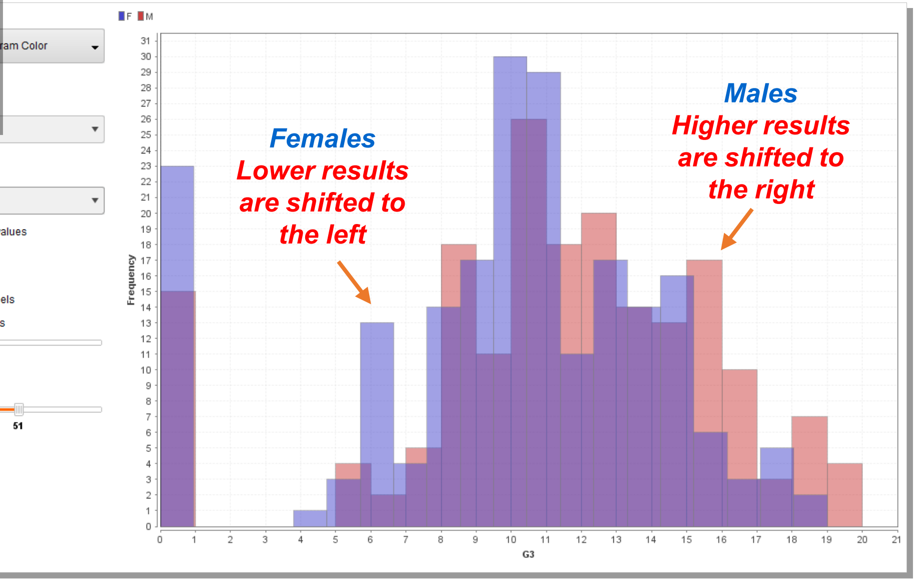
☐ Rotate labels

Number of Bins 20

Opaqueness 51

- Two variables can also be compared by overlaying their histograms (left).
- Overlaid histograms of student results in assessment G2 (blue) and G3 (orange) demonstrate similar distribution of marks, with many more G3 grades exceeding those in G2 (shift to the right).
- At the same time the final G3 results include more failures than those given in G2 (possibly due to course incompletion).

- It is also possible to overlay histograms of two variable subsets, e.g. split by the values of another nominal variable (right).
- Overlaid histograms of G3 results, as attained by female (blue) and male (orange) students, indicate that while the overall distributions of results are similar, males receive higher grades in Mathematics than females.



...ram Color

G3

Color: sex

☐ Absolute values

☐ Log Scale

☐ Rotate labels

Number of Bins 20

Opaqueness 51

A Self-Help Guide to RapidMiner Studio

Lecture demonstration using the “student-mat.csv” data set

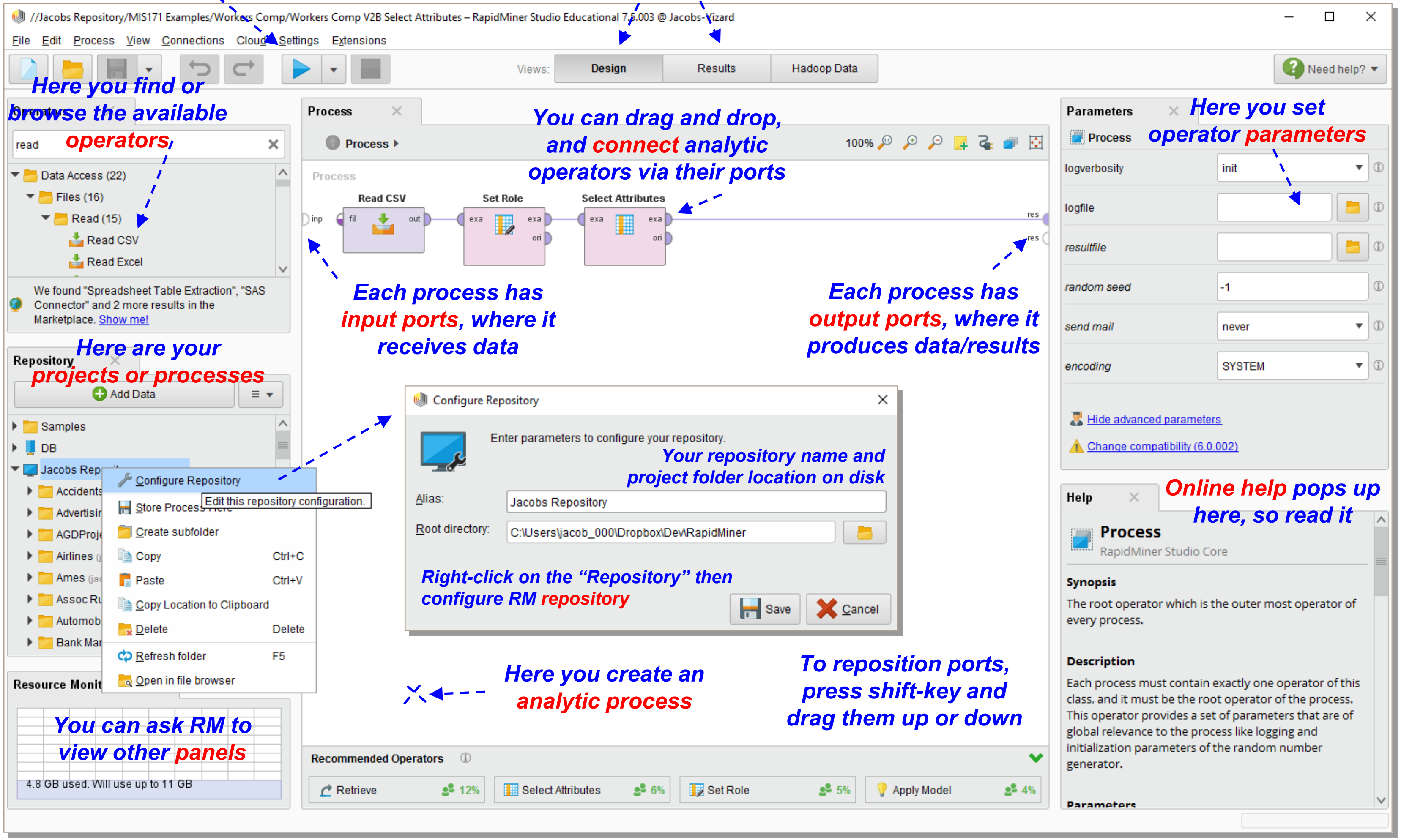
RM in 10 Easy Steps

1) *Install RapidMiner*
2) Create a project folder on your H: Drive
3) Create a data folder inside the project folder
4) Place CSV files in the data folder
5) Start RapidMiner
6) Configure RM repository to point to your Project folder and be named as you like
7) Create a RM process
8) Save it in the project folder (or sub-folder)
9) Run the process
10) Explore the results!

DEAKIN BUSINESS SCHOOL

Press Run to execute the process

RM has two (or more) important views: Design View in which you design your analytic process and Results View in which you can inspect data produced by your analytics process, in tables and charts.



//Jacobs Repository/MIS171 Examples/Workers Comp/Workers Comp V2B Select Attributes – RapidMiner Studio Educational 7.5.003 @ Jacobs-Wizard

Operators search: read

- Data Access (22)
  - Files (16)
    - Read (15)
      - Read CSV
      - Read Excel

We found "Spreadsheet Table Extraction", "SAS Connector" and 2 more results in the Marketplace. Show me!

Repository:

- Samples
- DB
- Jacobs Rep...
  - Accidents
  - Advertisin
  - AGDProj
  - Airlines
  - Ames
  - Assoc Ru
  - Automob
  - Bank Mar

Configure Repository	
Store Process Here	
Create subfolder	
Copy	Ctrl+C
Paste	Ctrl+V
Copy Location to Clipboard	
Delete	Delete
Refresh folder	F5
Open in file browser	

Edit this repository configuration.

Resource Monitor: 4.8 GB used. Will use up to 11 GB

Parameters:

logverbosity	init
logfile	
resultfile	
random seed	-1
send mail	never
encoding	SYSTEM

Hide advanced parameters

Change compatibility (6.0.002)

Help:

Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization parameters of the random number generator.

Parameters

Configure Repository

Enter parameters to configure your repository.

Alias: Jacobs Repository

Root directory: C:\Users\jacob_000\Dropbox\Dev\RapidMiner

Save | Cancel

Recommended Operators: Retrieve 12% | Select Attributes 6% | Set Role 5% | Apply Model 4%

Here you find or browse the available operators

You can drag and drop, and connect analytic operators via their ports

Here you set operator parameters

Each process has input ports, where it receives data

Each process has output ports, where it produces data/results

Here are your projects or processes

Your repository name and project folder location on disk

Right-click on the "Repository" then configure RM repository

Online help pops up here, so read it

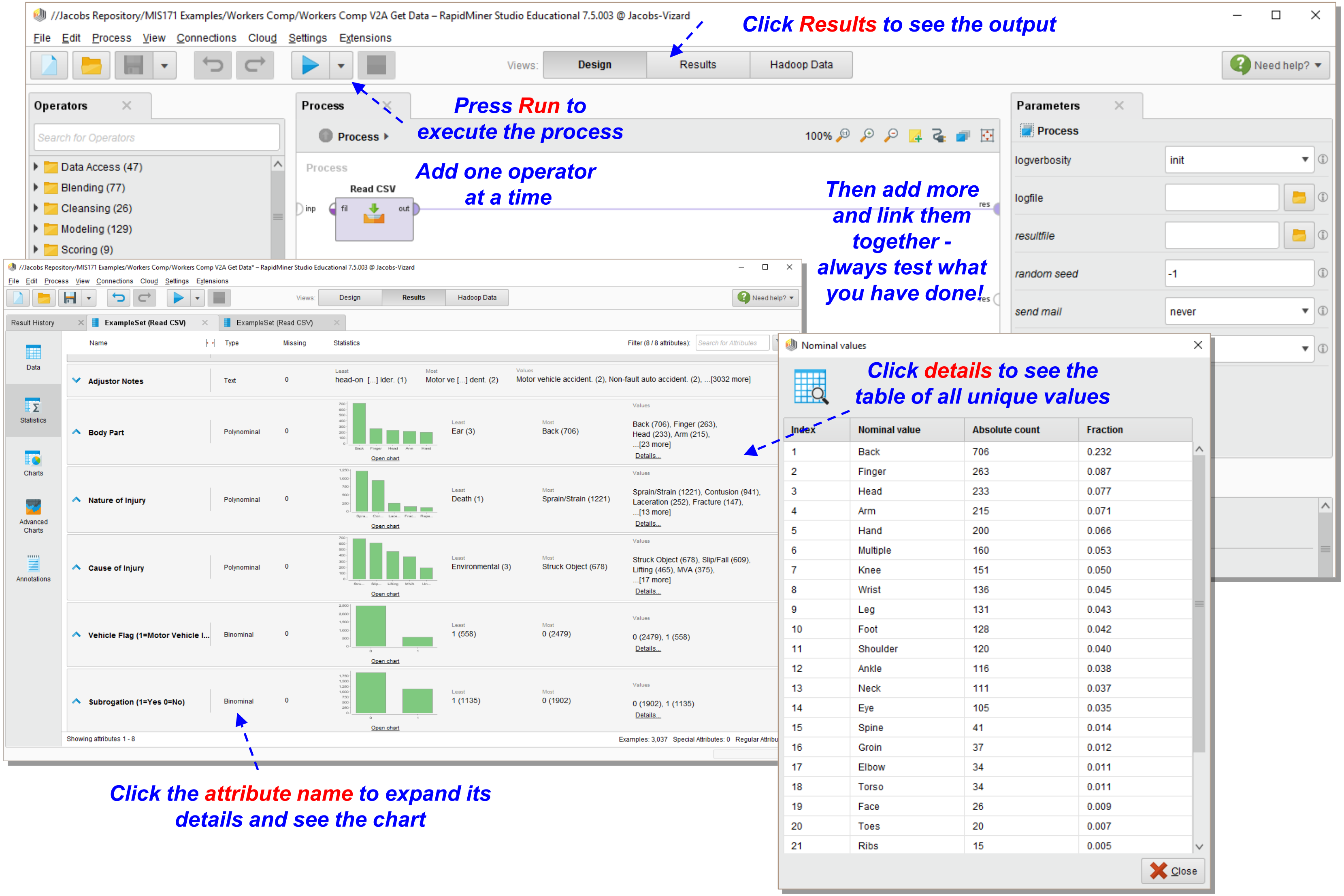
Here you create an analytic process

To reposition ports, press shift-key and drag them up or down

You can ask RM to view other panels

RM in Step 7 a, b, c

- Never develop your project in one monumental leap!
- Work in small steps, so execute this “mini” model first and see the results.
- Even the simplest process, consisting of reading the data in and returning it for further examination is valuable.
- Check the basic stats (min, max, mean, median, mode), distribution and values of all attributes.
- Only then add one more operator at a time and inspect its result.
- Do not forget to save your work often.



Click Results to see the output

Press Run to execute the process

Add one operator at a time

Then add more and link them together - always test what you have done!

Click details to see the table of all unique values

Click the attribute name to expand its details and see the chart

//Jacobs Repository/MIS171 Examples/Workers Comp/Workers Comp V2A Get Data – RapidMiner Studio Educational 7.5.003 @ Jacobs-Vizard

Process	
logverbosity	init
logfile	
resultfile	
random seed	-1
send mail	never

//Jacobs Repository/MIS171 Examples/Workers Comp/Workers Comp V2A Get Data* – RapidMiner Studio Educational 7.5.003 @ Jacobs-Vizard

Filter (8 / 8 attributes):

Name	Type	Missing	Least	Most	Values
Adjustor Notes	Text	0	head-on [...] lder. (1)	Motor ve [...] dent. (2)	Motor vehicle accident. (2), Non-fault auto accident. (2), ...[3032 more]
Body Part	Polynominal	0	Ear (3)	Back (706)	Back (706), Finger (263), Head (233), Arm (215), ...[23 more]
Nature of Injury	Polynominal	0	Death (1)	Sprain/Strain (1221)	Sprain/Strain (1221), Contusion (941), Laceration (252), Fracture (147), ...[13 more]
Cause of Injury	Polynominal	0	Environmental (3)	Struck Object (678)	Struck Object (678), Slip/Fall (609), Lifting (465), MVA (375), ...[17 more]
Vehicle Flag (1=Motor Vehicle I...	Binominal	0	1 (558)	0 (2479)	0 (2479), 1 (558)
Subrogation (1=Yes 0=No)	Binominal	0	1 (1135)	0 (1902)	0 (1902), 1 (1135)

Showing attributes 1 - 8

Examples: 3,037 Special Attributes: 0 Regular Attribu

Nominal values

Index	Nominal value	Absolute count	Fraction
1	Back	706	0.232
2	Finger	263	0.087
3	Head	233	0.077
4	Arm	215	0.071
5	Hand	200	0.066
6	Multiple	160	0.053
7	Knee	151	0.050
8	Wrist	136	0.045
9	Leg	131	0.043
10	Foot	128	0.042
11	Shoulder	120	0.040
12	Ankle	116	0.038
13	Neck	111	0.037
14	Eye	105	0.035
15	Spine	41	0.014
16	Groin	37	0.012
17	Elbow	34	0.011
18	Torso	34	0.011
19	Face	26	0.009
20	Toes	20	0.007
21	Ribs	15	0.005

Summary and Questions for Review

- What is predictive analytics?
- What are its main applications?
- What areas of knowledge contribute to its methods and techniques?
- What is the role of modelling in data analytics?
- What is the difference between exploratory, decision and predictive models?
- What are the similarities and differences between different kinds of predictions?
- Why data mining is important in predictive analytics?
- What statistical, machine learning and text analytics methods are used in predictive analytics?
- What open source and commercial tools are used for data mining and prediction?

- What is an analytic process?
- What are its main steps?
- Why is it important to study properties of observed attributes?
- Why bother studying attribute relationships?
- Name 5 statistics that could assist gaining insights into collected data.
- What are the benefits and pitfalls of data visualisation?
- Demo: What is the role of label and predictor attributes? What are they called in statistics?
- Demo: What statistics can be used to assess the model performance?
- Demo: Explain the main steps that are taken in the investigation of model performance.
- Demo: Explain the main principles, benefits and pitfalls of holdout validation.