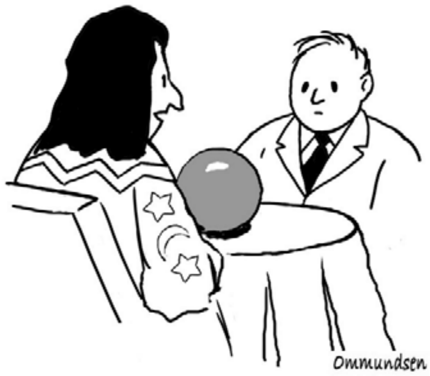
Master of Technology in Knowledge Engineering

Bayesian Modeling Workshop

Dr. Barry Shepherd
Institute of Systems Science
National University of Singapore
Email: barryshepherd@nus.edu.sg



"Is this needed for a Bayesian analysis?"

Workshop Goals

We use **US vehicle crash** data obtained from the Bayesia website* – see the Bayesia case study pdf file for detailed description of the problem domain

Goals

- Build and compare Bayesian Network prediction models to predict the likely injury level for vehicle occupants
- Interact with the *GeNIe* Bayesian Net to understand what factors impact vehicle safely (as indicated by OA-MAIS)

See on IVLE for this module:
vehicle_safety_NASS2010_2000_2012.csv
vehicle_safety_v20b.pdf

Variable Name	Long Name	Units/States	Comment
GV_CURBWGT	Vehicle Curb Weight	kg	
GV_DVLAT	Lateral Component of Delta V	km/h	
GV_DVLONG	Longitudinal Component of Delta V	km/h	
GV_ENERGY	Energy Absorption	J	
GV_FOOTPRINT	Vehicle Footprint	m^2	calculated as WHEELBAS x ORIGAVTW
GV_LANES	Number of Lanes	count	
GV_MODELYR	Vehicle Model Year	year	
GV_OTVEHWGT	Weight Of The Other Vehicle	kg	
GV_SPLIMIT	SpeedLimit	mph	converted into U.S. customary units
GV_WGTCDTR	Truck Weight Code	missing = Passenger Vehicle	
		6,000 and less	
		6,001 - 10,000	
OA_AGE	Age of Occupant	years	
OA_BAGDEPLY	Air Bag System Deployed	Nondeployed	
		Bag Deployed	
OA_HEIGHT	Height of Occupant	cm	
OA_MAIS	Maximum Known Occupant AIS	Not Injured	AIS Probability of Death
		Minor Injury	0%
		Moderate Injury	1-2%
		Serious Injury	8-10%
		Severe Injury	5-50%
		Critical Injury	5-50%
		Maximum Injury	100% (Unsurvivable)
		Unknown	Missing Value
OA_MANUSE	Manual Belt System Use	Used	
		Not Used	
OA_SEX	Occupant's Sex	Male	
		Female	
OA_WEIGHT	Occupant's Weight	kg	
VE_GAD1	Deformation Location (Highest)	Left	
		Front	
		Rear	
		Right	
VE_PDOF_TR	Clock Direction for Principal Direction of Force (Highest)	Degrees	Transformed variable, rotated 135 degrees counterclockwise

Workshop Goals

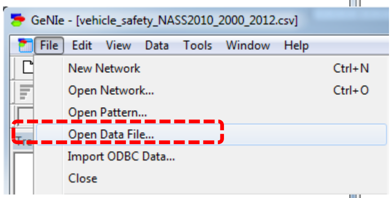
- Build a Naïve Bayes network + one other Bayesian network and then compare their results (e.g. compare prediction accuracy)
- Use any tool(s): GeNIe, SPSS Modeler, BayesiaLab, JMP, R
  - GeNIe ~ Naïve Bayes, Tree Augmented Naïve Bayes (TAN) ** easy for beginners
  - SPSS Modeler ~ TAN, Markov Blanket
  - JMP ~ Naïve Bayes
  - BayesiaLab ~ Naïve Bayes, Markov Blanket (will need to self-learn)
  - R ~ lots of libraries available, e.g. bnlearn (NaïveBayes, TAN), e1071 (NaiveBayes)
- If you compare across tools then try to ensure you use the same dataset and training/test set division with all tools (else comparing results isn't accurate)
  - GeNIe only works with discrete variables hence you must perform discretization of numerical fields first (e.g. using binning) – can be done in GeNIe (or Excel)
  - SPSS Modeler automatically bins continuous variables (also has nice binning tool)
  - JMP assumes continuous variables are Normally distributed

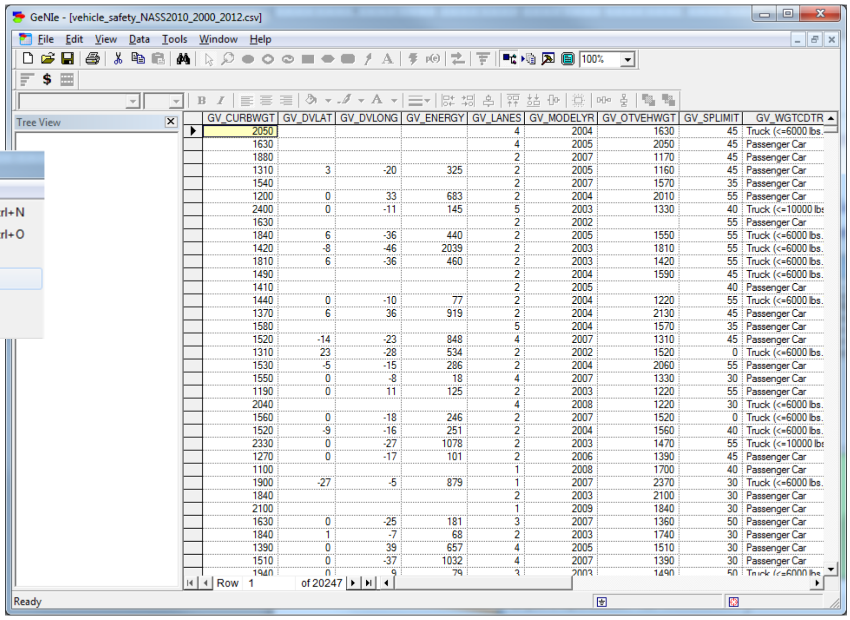
Tool Review – a quick look at relevant features

- GeNIe — https://download.bayesfusion.com/files.html?category=Academia
- SPSS Modeler — https://nus.onthehub.com
- JMP — see IVLE for instructions

GeNIe: Loading the Data

- Load the data using: ***File->Open Data File***
- Networks can ***only*** be built from categorical columns – you must bin all numerical variables first

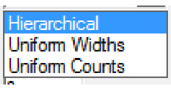


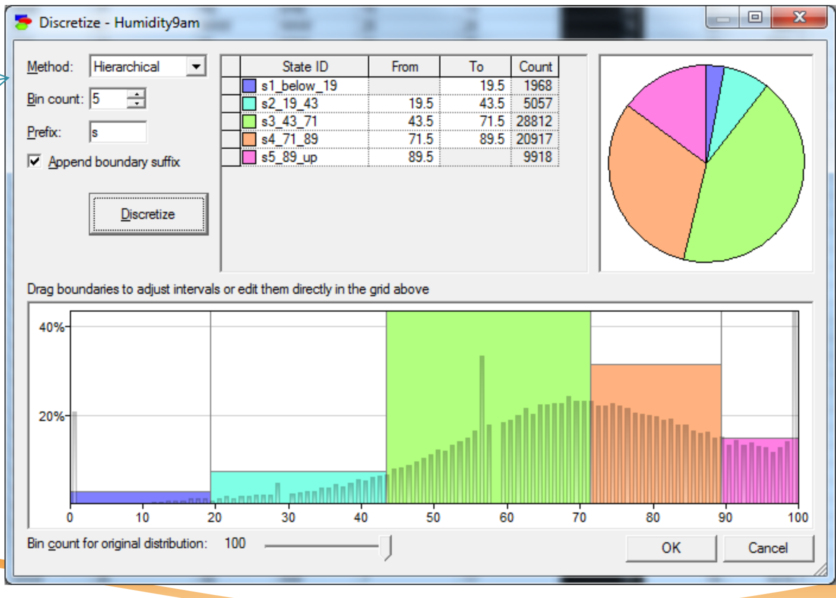


GeNIe: Discretizing variables

- Has a nice graphical binning tool
- Best to bin one variable at a time (even though multiple columns can be selected)
- Select a numeric column from the data on display, then select: ***Data->Discretize*** (the numeric values will be overwritten with the discretized ones)

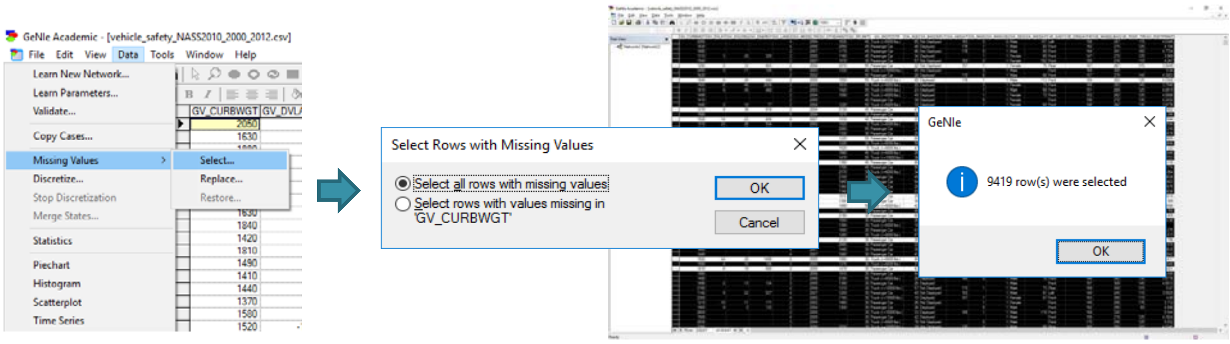
Other methods



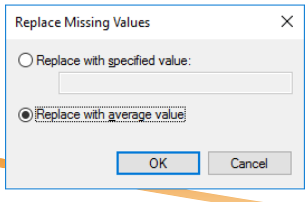


GeNIe: Handling Missing Values

- Naïve Bayes works with missing values, but for other network architectures better results might be obtained by filling in the missing values with estimates. For this workshop, you may use ***any method or tool*** to handle the missing values – you choose!

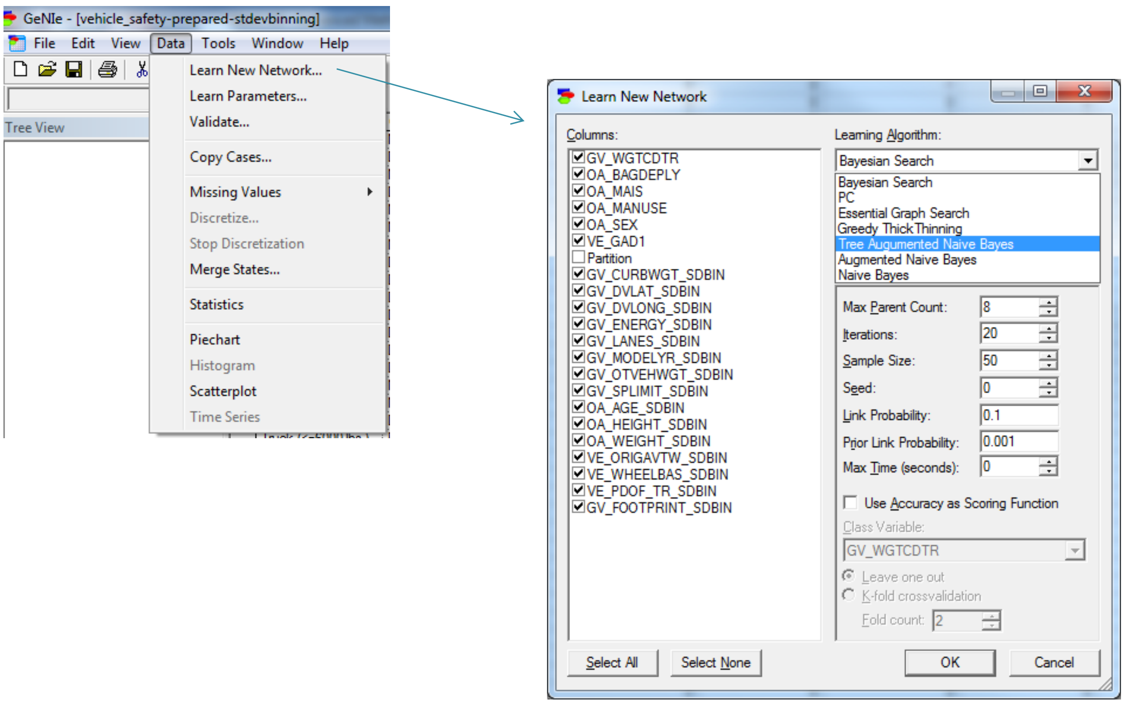


- There are 9,419 rows with missing values out of 20,240 records – nearly 50%, too many to ignore (i.e. to delete)!
- Fortunately only 43 rows have a missing target value



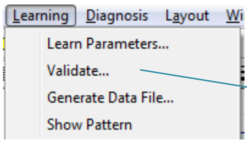
- Replacing with averages is quick but crude, but can be effective
- Building models to impute them can be more accurate (but not mandatory for this workshop)

GeNIe: Learning Networks



GeNIe: Validating the Learned Network

Test the network using Validate…

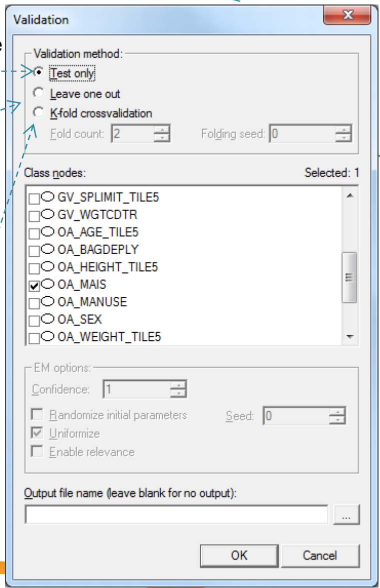


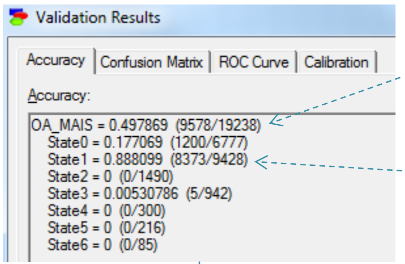
Just click OK on this screen

Apply existing network to whole data set

Leave each example out in turn and build a fresh network, then test on the left-out example

Build and test N networks (N = #folds)

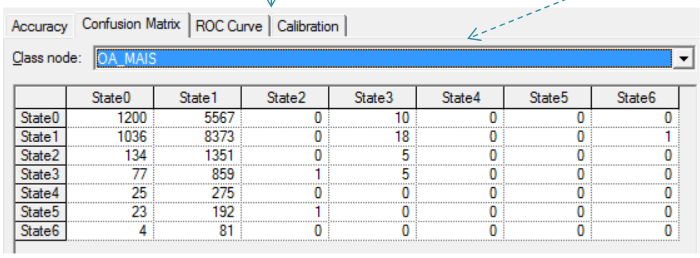




Overall accuracy = sum of diagonal/total rows

Per class accuracy = #correct/row sum

Rows = actuals Columns = predictions



Accuracy | Confusion Matrix | ROC Curve | Calibration

Class node: OA_MAIS

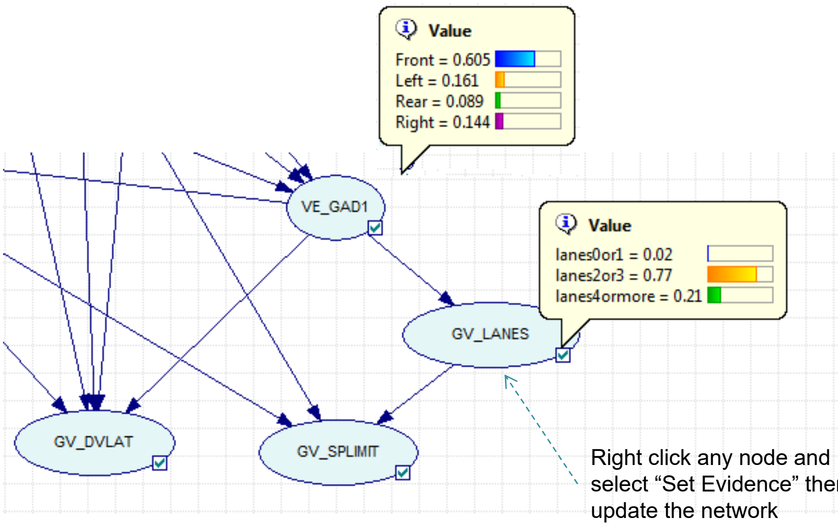
	State0	State1	State2	State3	State4	State5	State6
State0	1200	5567	0	10	0	0	0
State1	1036	8373	0	18	0	0	1
State2	134	1351	0	5	0	0	0
State3	77	859	1	5	0	0	0
State4	25	275	0	0	0	0	0
State5	23	192	1	0	0	0	0
State6	4	81	0	0	0	0	0

These add up to 19,238 (total rows in data) if "Test only" was selected

Exploratory Analysis in GeNIe

- It is possible to set evidence for specific node(s) and see the impact on the immediate neighbour nodes (and the impact on all nodes, but impact may be small for far away nodes)
- E.g. to see the impact of high speed, multi-lane highways on vehicle impact zone (VE_GAD) we set GV_LANES to be >= 4 and speed-limit to high (exact settings will depend on the degree of binning you performed)

For exploring interactions also try using unsupervised learning to build the net (e.g. "Bayesian search" or "Greedy Thick Thinning" learning algorithms)

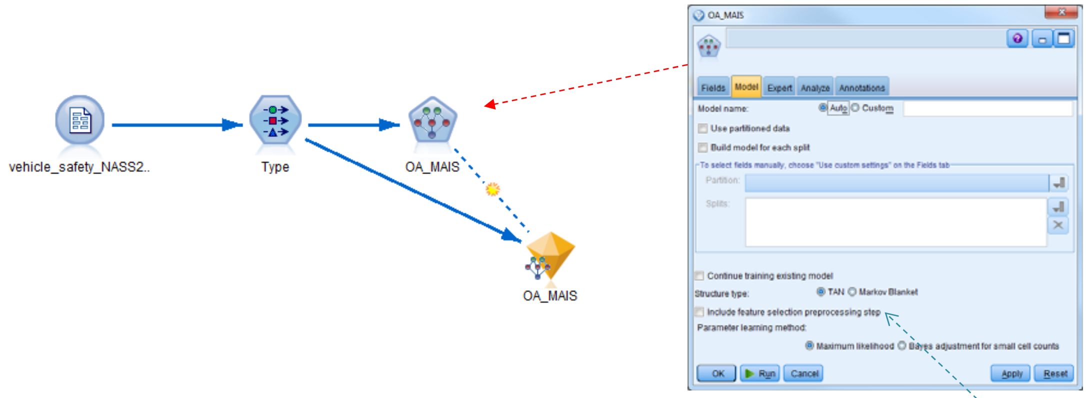


Right click any node and select "Set Evidence" then update the network

SPSS Modeller: Bayes Net Node

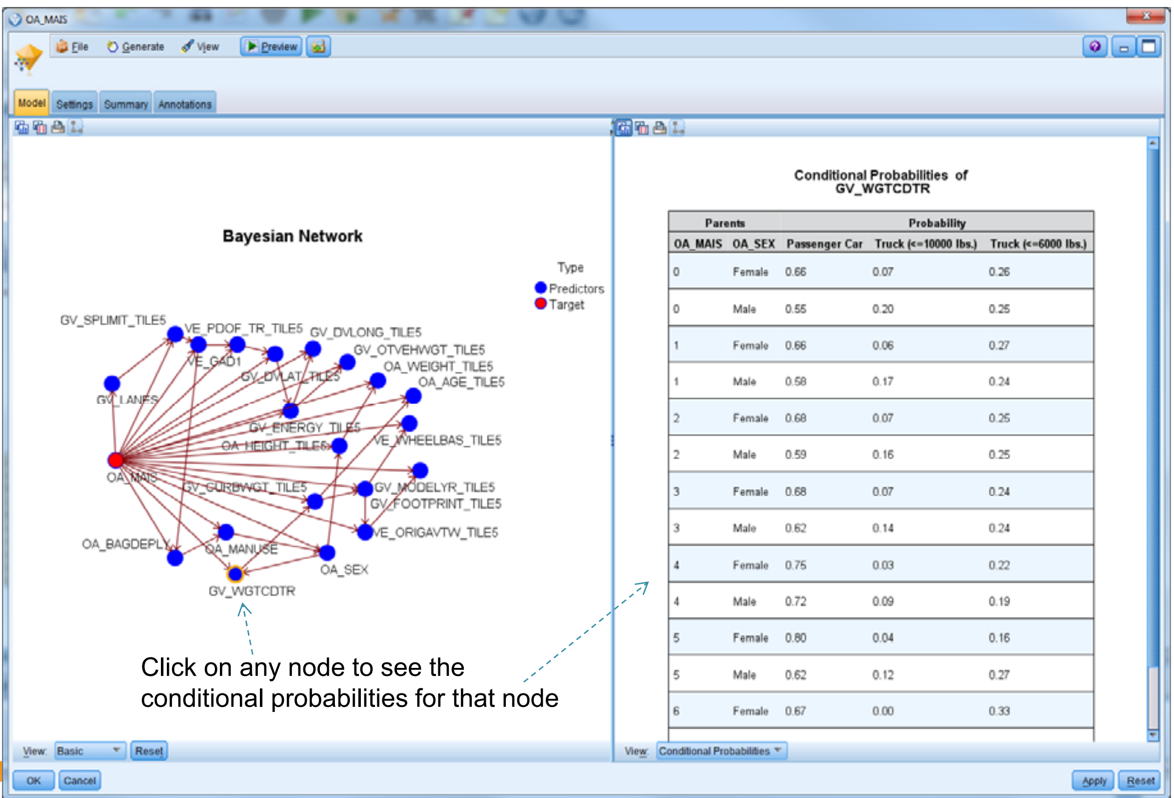
Bayes Net

- Use like any other modeling node
- Target fields must be *Nominal*, *Ordinal*, or *Flag*.
- Inputs can be fields of any type. Continuous (numeric range) input fields will be automatically binned; however, if the distribution is skewed, you may obtain better results by manually binning the fields using a Binning node before the Bayesian Network node.



Two network structures are available

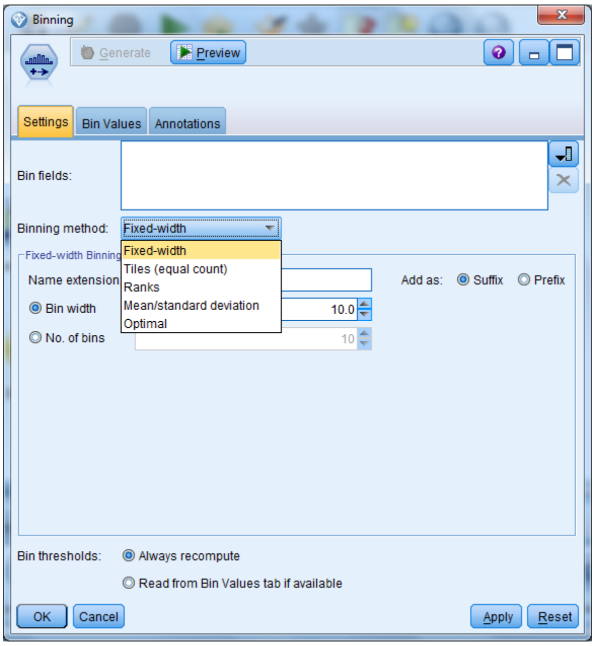
SPSS: Examining the Built Network



SPSS: Binning Node

- The Bayes Node will automatically bin numerical inputs
- But its often better if you do it explicitly yourself using the Binning node…

  - Fixed-width binning
  - Tiles (equal count or sum)
  - Mean and standard deviation
  - Ranks
  - Optimized relative to a categorical "supervisor" field



SPSS : Binning Node

- Fixed Width Binning – you specify the width of the bin (integer or real). The default with is 10, for example:

Table 1. Bins for Age with range 18–65

Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
>=13 to <23	>=23 to <33	>=33 to <43	>=43 to <53	>=53 to <63	>=63 to <73

- Tiles ~ Fixed size bins - can be based on record count or sum of values.

  Options:
  - ***Quartile.*** *Generate 4 bins, each containing 25% of the cases.*
  - ***Quintile.*** *Generate 5 bins, each containing 20% of the cases.*
  - ***Decile.*** *Generate 10 bins, each containing 10% of the cases.*
  - ***Vingtile.*** *Generate 20 bins, each containing 5% of the cases.*
  - ***Percentile.*** *Generate 100 bins, each containing 1% of the cases.*
  - ***Custom N.*** *Select to specify the number of bins.*

SPSS Modeler Binning Node

- Mean and Standard Deviation

- **+/– 1 standard deviation.** Select to generate three bins.
- **+/– 2 standard deviations.** Select to generate five bins.
- **+/– 3 standard deviations.** Select to generate seven bins.

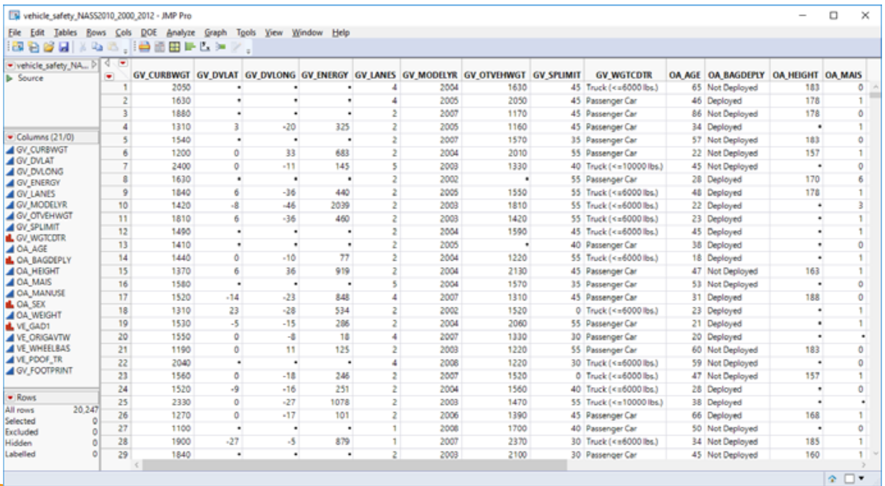
For example, selecting +/–1 standard deviation results in the three bins as calculated and shown in the following table.

Table 1. Standard deviation bin example

Bin 1	Bin 2	Bin 3
`x < (Mean - Std. Dev)`	`(Mean - Std. Dev) <= x <= (Mean + Std. Dev)`	`x > (Mean + Std. Dev)`

JMP

- After loading the data….
- Select: *Cols->Column Info..* to change data types, e.g. convert OS_MAIS into nominal (or ordinal)
- Select: *Analyze->Predictive Model->Make ValidationColumn* to specify train/testset
- Select: *Analyze->Predictive Model->NaiveBayes* to build the network



	GV_CURBWGT	GV_DVLAT	GV_DVLONG	GV_ENERGY	GV_LANES	GV_MODELYR	GV_OTVEHWGT	GV_SPLIMIT	GV_WGTCDTR	OA_AGE	OA_BAGDEPLY	OA_HEIGHT	OA_MAIS
1	2050	•	•	•	4	2004	1630	45	Truck (<=6000 lbs.)	65	Not Deployed	183	0
2	1630	•	•	•	4	2005	2050	45	Passenger Car	46	Deployed	178	1
3	1880	•	•	•	2	2007	1170	45	Passenger Car	86	Not Deployed	178	0
4	1310	3	-20	325	2	2005	1160	45	Passenger Car	34	Deployed	•	1
5	1540	•	•	•	2	2007	1570	35	Passenger Car	57	Not Deployed	183	0
6	1200	0	33	683	2	2004	2010	55	Passenger Car	22	Not Deployed	157	1
7	2400	0	-11	145	5	2003	1330	40	Truck (<=10000 lbs.)	45	Not Deployed	•	0
8	1630	•	•	•	2	2002	•	55	Passenger Car	28	Deployed	170	6
9	1840	6	-36	440	2	2005	1550	55	Truck (<=6000 lbs.)	48	Deployed	178	1
10	1420	-8	-46	2039	2	2003	1810	55	Truck (<=6000 lbs.)	22	Deployed	•	3
11	1810	6	-36	460	2	2003	1420	55	Truck (<=6000 lbs.)	23	Deployed	•	1
12	1490	•	•	•	2	2004	1590	45	Truck (<=6000 lbs.)	45	Deployed	•	1
13	1410	•	•	•	2	2005	•	40	Passenger Car	38	Deployed	•	0
14	1440	0	-10	77	2	2004	1220	55	Truck (<=6000 lbs.)	18	Deployed	•	1
15	1370	6	36	919	2	2004	2130	45	Passenger Car	47	Not Deployed	163	1
16	1580	•	•	•	5	2004	1570	35	Passenger Car	53	Not Deployed	•	0
17	1520	-14	-23	848	4	2007	1310	45	Passenger Car	31	Deployed	188	0
18	1310	23	-28	534	2	2002	1520	0	Truck (<=6000 lbs.)	23	Deployed	•	1
19	1530	-5	-15	286	2	2004	2060	55	Passenger Car	21	Deployed	•	1
20	1550	0	-8	18	4	2007	1330	30	Passenger Car	20	Deployed	•	•
21	1190	0	11	125	2	2003	1220	55	Passenger Car	60	Not Deployed	183	0
22	2040	•	•	•	4	2008	1220	30	Truck (<=6000 lbs.)	59	Not Deployed	•	0
23	1560	0	-18	246	2	2007	1520	0	Truck (<=6000 lbs.)	47	Not Deployed	157	1
24	1520	-9	-16	251	2	2004	1560	40	Truck (<=6000 lbs.)	28	Deployed	•	0
25	2330	0	-27	1078	2	2003	1470	55	Truck (<=10000 lbs.)	38	Deployed	•	•
26	1270	0	-17	101	2	2006	1390	45	Passenger Car	66	Deployed	168	1
27	1100	•	•	•	1	2008	1700	40	Passenger Car	50	Not Deployed	•	0
28	1900	-27	-5	879	1	2007	2370	30	Truck (<=6000 lbs.)	34	Not Deployed	185	1
29	1840	•	•	•	2	2003	2100	30	Passenger Car	45	Not Deployed	160	1

What to Hand In/Upload to IVLE

- Your code + any updated data file (e.g. after binning) – ZIP them together
- A short report telling me what you did and the results you obtained
- Report should include:
  1. Data Preparation
     - Describe what pre-processing you did
     - Describe how you split the data in train & test sets
  2. Models Built
     - Paste/draw a pic of the networks
     - Any other useful details?
  3. Results
     - Ideally a confusion matrix and prediction accuracy for each model

This assignment counts for 10 marks

Hand-in by March 22nd