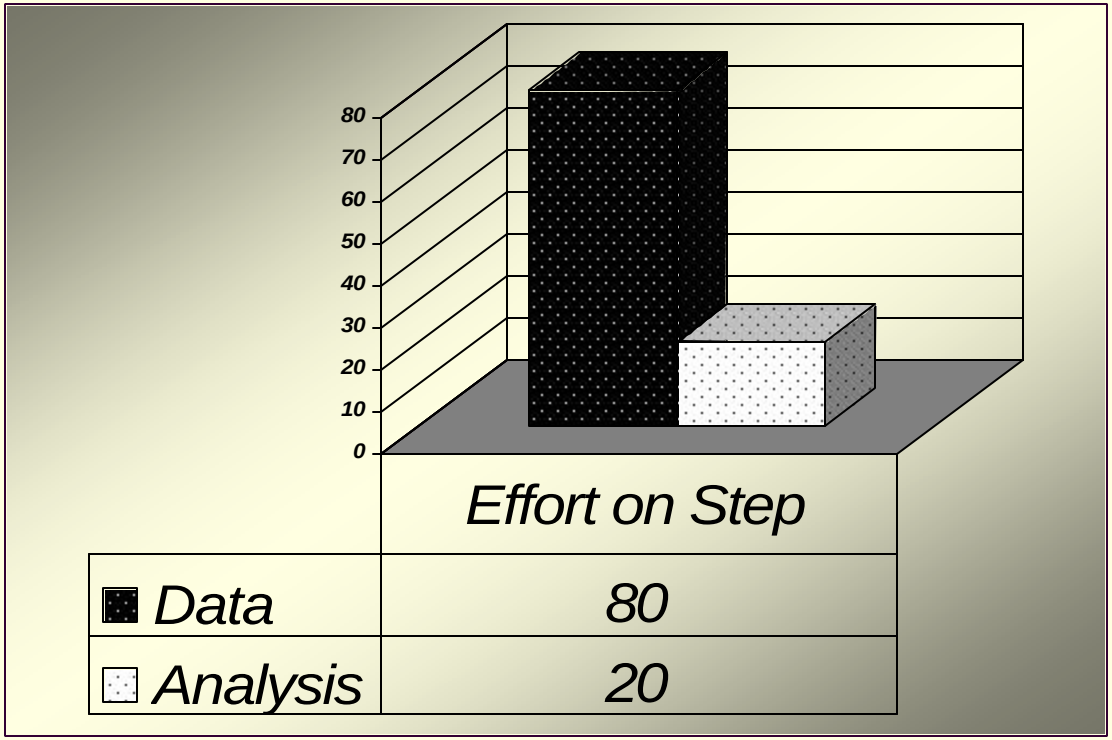
Data: The Undiscovered Country

WRG 2007 Predictive Modeling in Workers Compensation
Louise Francis, FCAS
Francis Analytics and Actuarial Data Mining, Inc.
www.data-mines.com
Louise_francis@msn.com

DATA

Data Preprocessing Effort



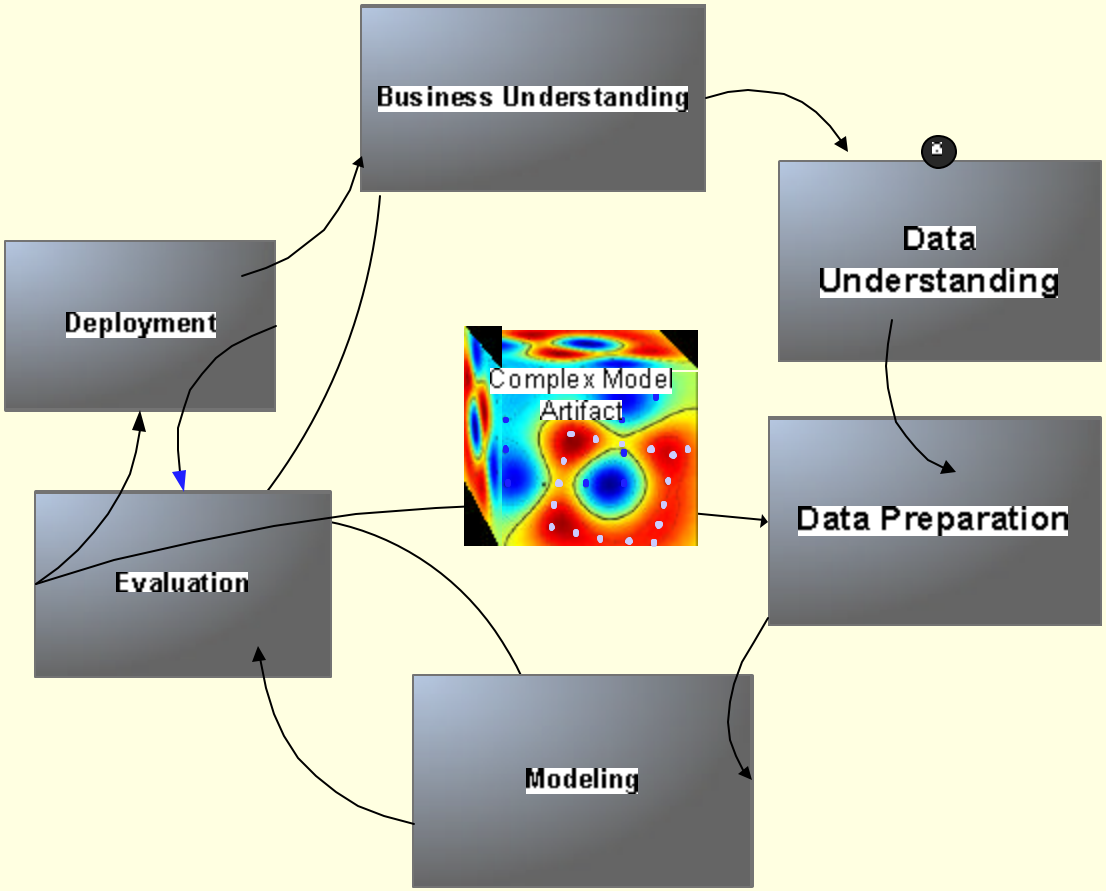
Objectives

- Introduce data preparation and where it fits in in modeling process
- Discuss Data Quality
- Focus on key steps in data preparation
  - Exploratory data analysis
    - Identify data glitches and errors
    - Understanding the data
    - Identify possible transformations
  - Data augmentation
  - Data reduction
  - What to do about missing data
  - Provide resources on data quality and data preparation

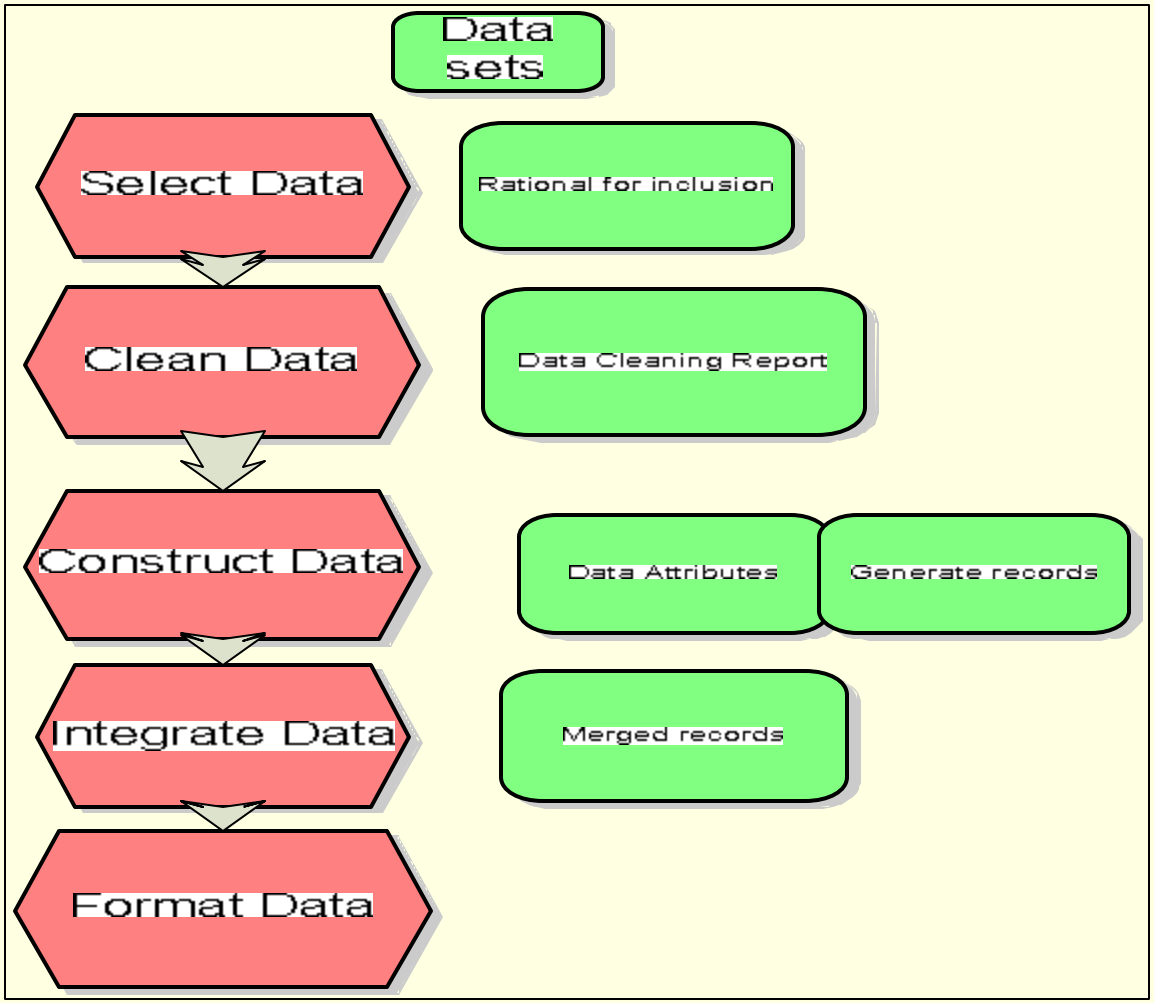
CRISP-DM

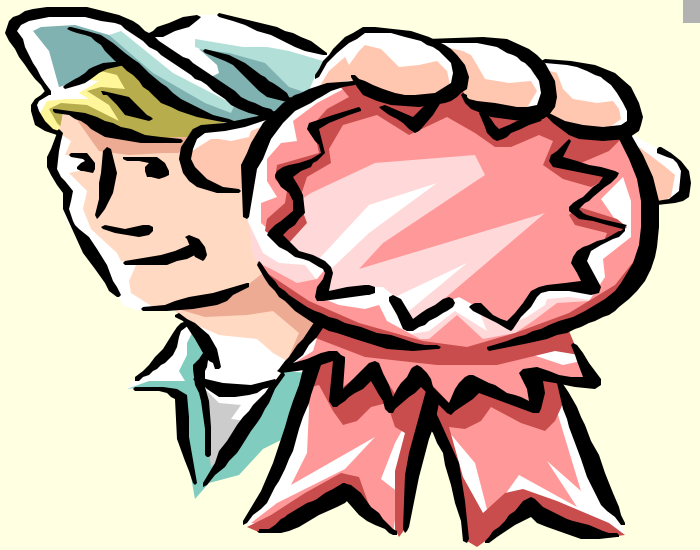
- Guidelines for data mining projects
- Gives overview of life cycle of data mining project
- Defines different phases and activities that take place in phase
- Source for information on CRISP-DM: www.spss.com

Modeling Process



Data Preprocessing

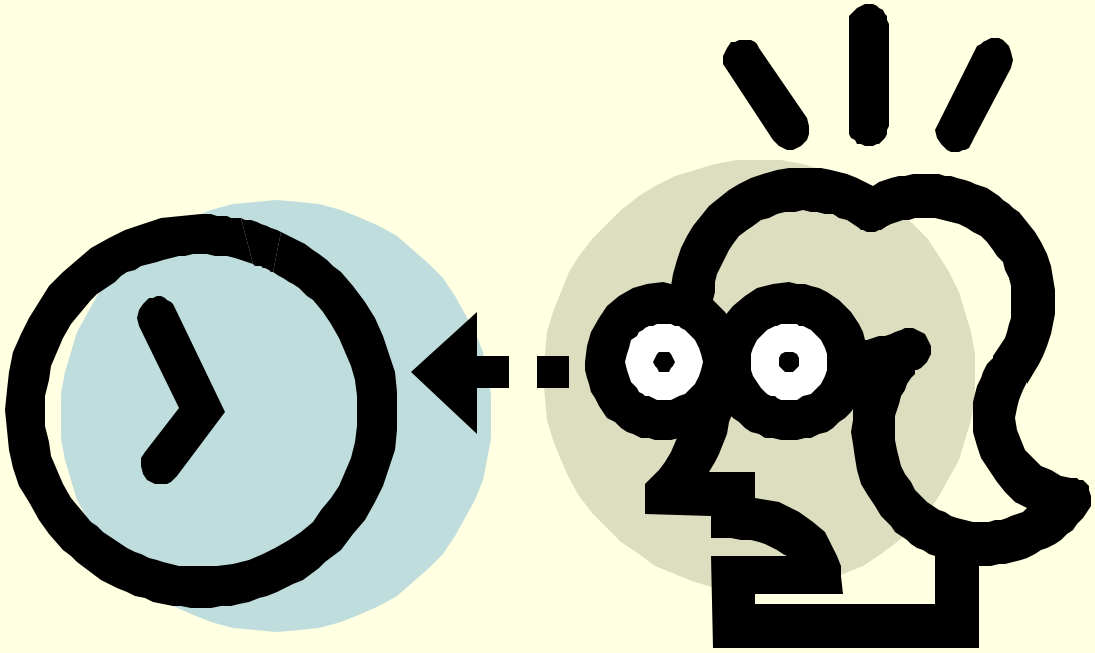




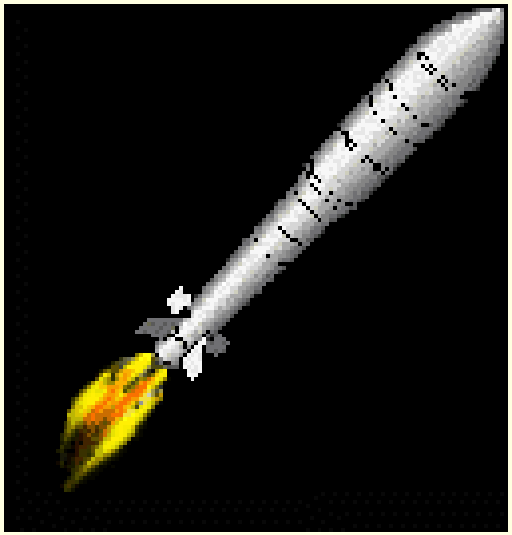
Data Quality Problem

Data Quality: A Problem

- Actuary reviewing a database



From Swamp Mud to Rocket Science



Data + Model = Results

It’s Not Just Us

- “In just about any organization, the state of information quality is at the same low level”
  - Olson, *Data Quality*

What is Data Quality?

Dimension	Conditions for high quality data
Representational consistency	Values for a particular attribute have the same representation across all tables (e.g. dates)
Organizational consistency	There is one organization-wide table for each entity and one organization-wide domain for each attribute
Row consistency	The values in a row are internally consistent (e.g. a home phone number's area code is consistent with a city's location)
Flexibility	The content and format of presentations can be readily altered to meet changing circumstances
Precision	Data values can be conveniently formatted to the required degree of accuracy
Granularity	Data are represented at the lowest level necessary to support all uses (e.g. hourly sales)

* From CAS White Paper on Information Quality

What is Data Quality?

Dimension	Conditions for high quality data
Stewardship	Responsibility has been assigned for managing data
Sharing	Data sharing is widespread across organizational units
Timeliness	A value's recentness matches the needs of the most time critical application requiring it. Values remain up to date.
Interpretation	Clients correctly interpret the meaning of data elements

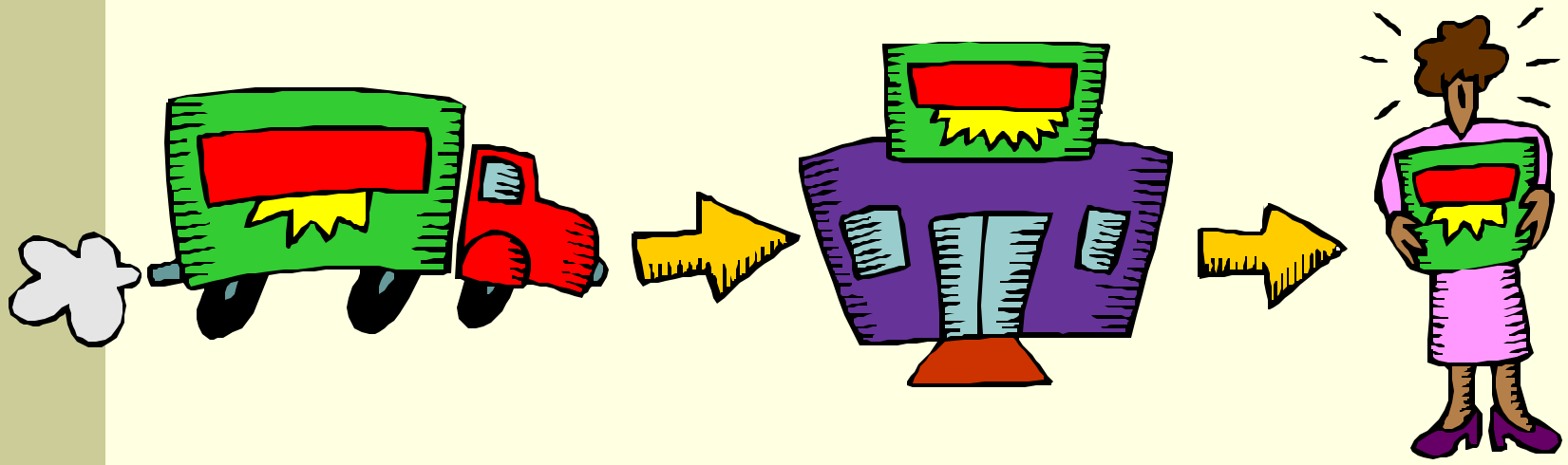
* From CAS White Paper on Information Quality

Data Quality Definition

- Quality for the intended purpose
- Data quality is relative

Some Consequences of poor data quality

- Affects quality (precision) of result
- Can’t do modeling project because of data problems
- If errors not found – modeling blunder



Data Preparation: Process

Data Preparation as a Process

- “There is no automatic tool that can be pointed at a data set and told to just “fix” the data” - p. 89 of *Data Preparation for Data Mining* by Dorian Pyle
- Inputs, Outputs, Models, and Decisions
- Survey the data
- Modeling tools and data preparation
- Stages of data preparation

Inputs, Outputs, Models, and Decisions

- Prepare the data
  - Training data
  - Testing data
  - A PIE-I (Prepared Information Environment Input Module)
  - A PIE-O (Prepared Information Output Module
    - Apply model to execution data
    - May require processing and transforming of variables

Identify data

- Internal
  - Policy
  - Claims
  - Billing Data
  - Provider bill review
  - Loss Control
  - HR
- External
  - Credit: D&B, Experian, Fair Isaac
  - Demographic: www.census.gov
  - Climate: http://www.melissadata.com/Lookups/ZipWeather.asp
  - BLS:http://research.stlouisfed.org/fred2/

Survey the Data

- High level overview of the data
  - exploratory data analysis (EDA)
    - Use statistical and graphical methods to understand structure and identify errors
  - audit data

Model the Data

- Model the data
  - Identify modeling procedure before assembling data
- Use the model
  - Identify data needed to deploy the model

Modeling Tools and Data Preparation

- Identify method
  - Neural networks
  - Generalized linear models
  - Trees
  - Clustering
- Identify software
  - Excel and Excel add-inns
  - Public domain such as R
  - Commercial software

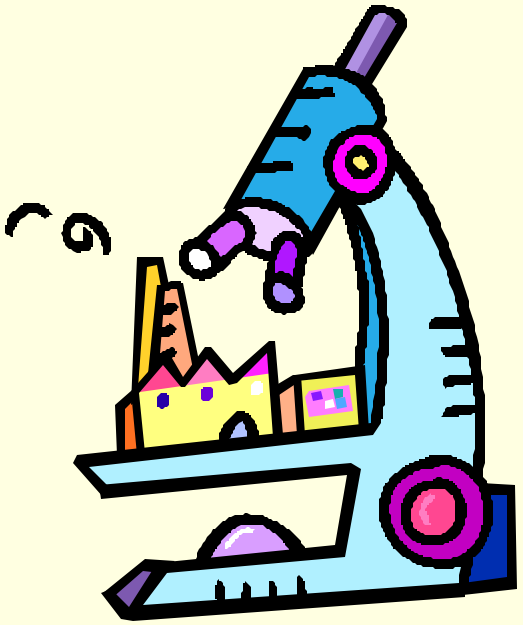
Modeling Tools and Data Preparation cont.

- What are data size limitations of software?
- What transformations are required by the software?
  - How are categorical variables coded?
    - Categories require a numeric code?
  - Max of 32 categories of categorical variables
- Other data requirements?

Modeling Tools: Missing Data

- How does software handle missing data?
  - Leave record out if data missing
  - Fill in a value, such as mean, for missing data
  - Give error message if any data on any variable is missing
  - This topic discussed in more detail later

Auditing the data



Basic Issues for auditing

- The source of supply
- The quantity of data
  - Number of records
  - Number of variables
- The quality of data
- Merging the databases

Auditing the data

- Examine a sample of records
  - Number of fields
  - Content of fields
  - Source of field
  - Type of values
  - What are typical values?
- Do you see evidence of glitches?
  - Such as: claimant birth date of 1890

Auditing data at the corporate level

- Used to assess and monitor quality of data
- Goal: To affect process generating the data
- Top down and bottom up approach
  - Top down – reconcile summary totals to other sources
  - Bottom up - sampling

Auditing data at the corporate level

- **Test the preparation of the data:** Measure the extent that data is correct, complete and timely
- **Test the data entry and data transfers:** Measure the extent that "all source documents and data records reach their appropriate destinations intact and in a timely manner."
- **Test the program controls:** A controlled processing environment will have procedures and checks to ensure that jobs are run in the right order, jobs are not accidentally run twice, total outputs equal total inputs, users are aware when programs end abnormally and so forth.
- **Test the output controls:** Measure the extent that all required output reports are accurate and "distributed to the appropriate individuals in a timely manner."
- **Test error procedures:** Measure the extent that the system detects and corrects errors in a timely manner.
  - From CAS paper on information quality

Auditing data

- Important goal:
  - What is feasibility of getting useful models with this data?

Inspect sample of data

AccDate	RptDate	SuitDate	SettlDate	Initial-Indemnityr-eserve	Initial-expense-reserve
6/23/1997	5/29/1997	5/1/1998	7/12/2001	20,000,000	10,000
6/23/1997	6/2/1998	4/11/1998	12/6/2001	1,000	-
10/7/1999	10/12/1999	7/16/2001	1/18/2003	7,500	1

Inspect sample of data, cont.

Finalexpenditureres erveQ8D Q8E	Attorney-InvolvementPlaintiff	AttorneyInvolvement-Insurer	Attorney-involvement-insured	Legal-stage-where-settlement-wasreached
525,000	Y	Y	N	4
150,000	Y	Y	N	3
327,871	P	Y	N	3
43,000	Y	Y	N	3

Data Exploration in Predictive Modeling

Exploratory Data Analysis

- Typically the first step in analyzing data
- Makes heavy use of graphical techniques
- Also makes use of simple descriptive statistics
- Purpose
  - Find outliers (and errors)
  - Explore structure of the data

Definition of EDA

Exploratory data analysis (EDA) is that part of statistical practice concerned with reviewing, communicating and using data where there is a low level of knowledge about its cause system.. Many **EDA** techniques have been adopted into data mining and are being taught to young students as a way to introduce them to statistical thinking.

- www.wikipedia.org

Example Data: Auto Liability

- Private passenger auto
- Some variables are:
  - Age
  - Gender
  - Marital status
  - Zip code
  - Earned premium
  - Number of claims
  - Incurred losses
  - Paid losses
  - Legal representation
  - Suspicion score (of fraud)

Example Data 2: Texas WC Claims Data

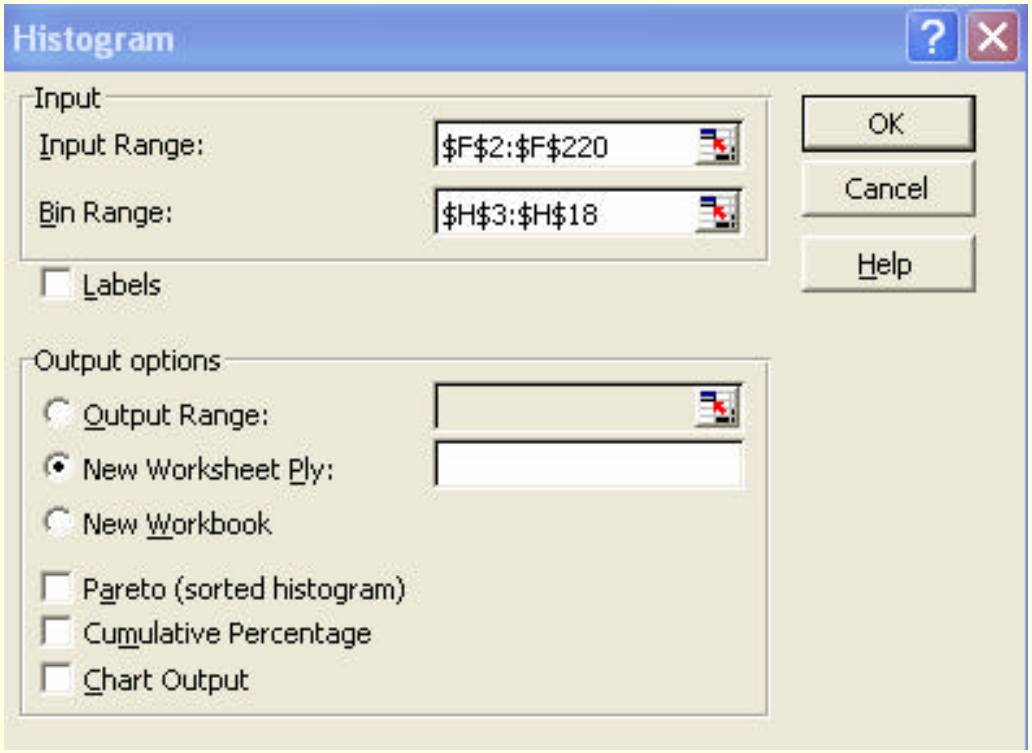
- Downloaded from Texas department of Insurance web site
- Closed claims > 25,000
  - Accident date, Report date, suit date, trial date, etc.
  - Initial indemnity reserves, final indemnity reserves, etc
  - Court verdict, amount of settlement
  - Attorney involvement, plaintiff, insurer
  - Injury code, cause of loss code

Some Methods for Numeric Data

- Visual
  - Histograms
  - Box and Whisker Plots
  - Stem and Leaf Plots
- Statistical
  - Descriptive statistics
  - Data spheres

Histograms

- Can do them in Microsoft Excel



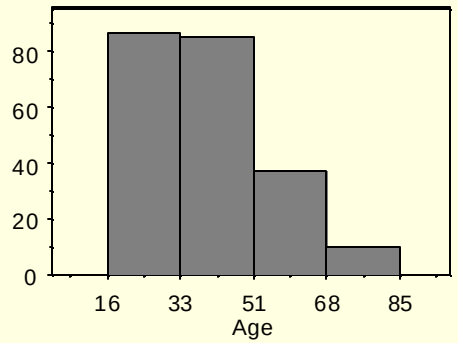
Histograms

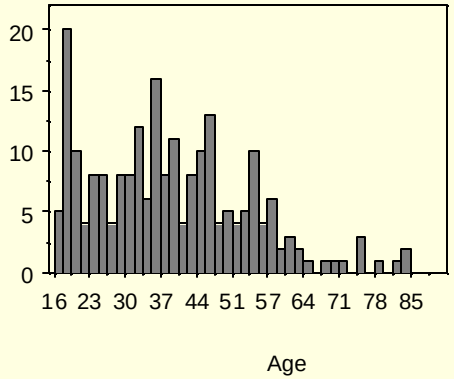
Frequencies for Age Variable

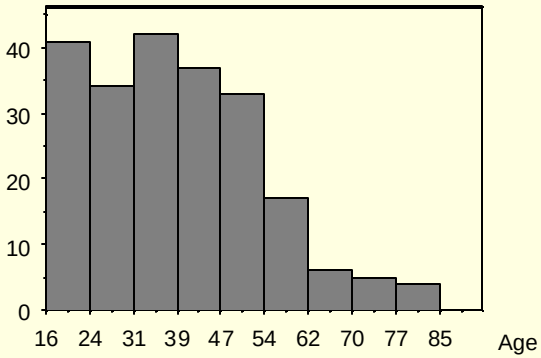
Bin	*Frequency*
20	2853
25	3709
30	4372
35	4366
40	4097
45	3588
50	2707
55	1831
60	1140
65	615
70	397
75	271
80	148
85	83
90	32
95	12
More	5

Histograms of Age Variable

Varying Window Size







Formula for Window Width

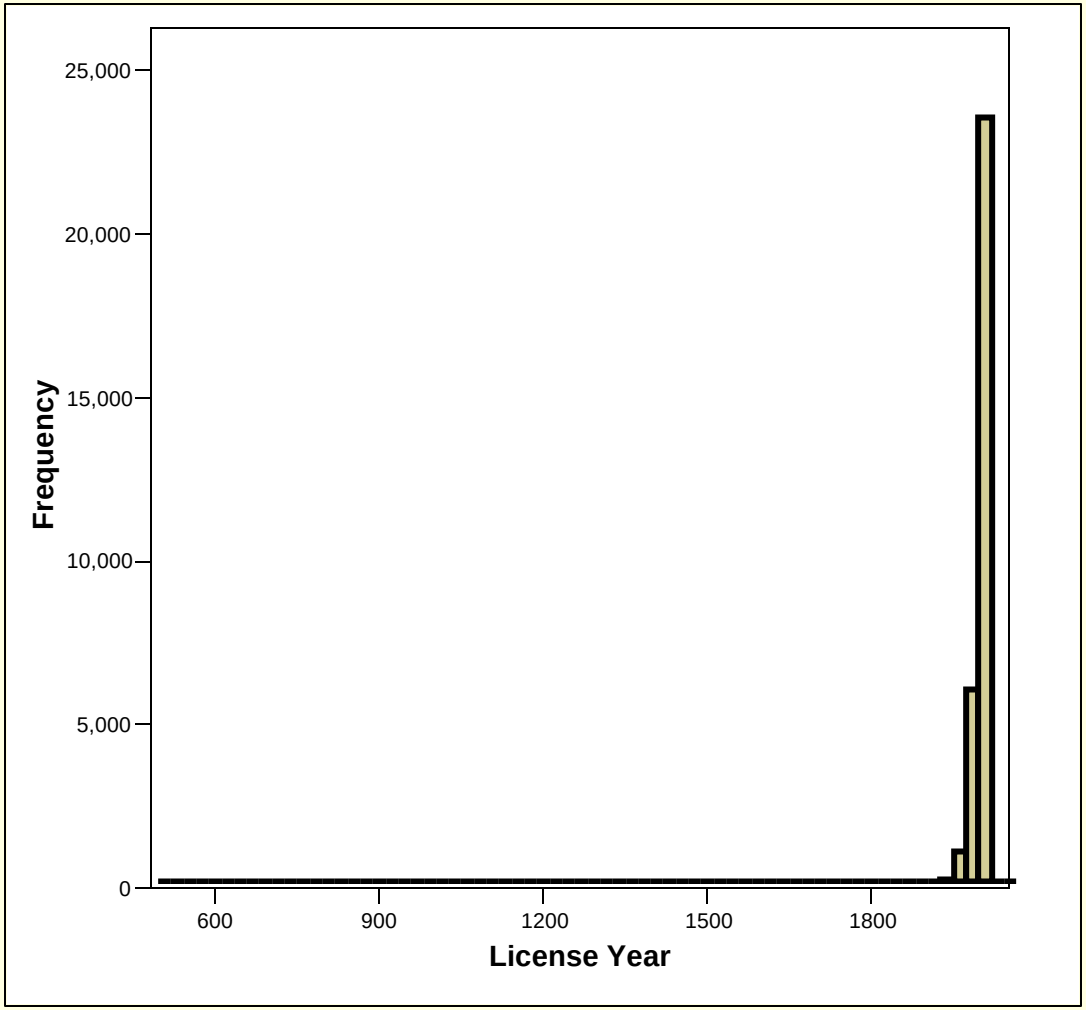
$$h = \frac{3.5s}{\sqrt[3]{N}}$$

s = standard deviation

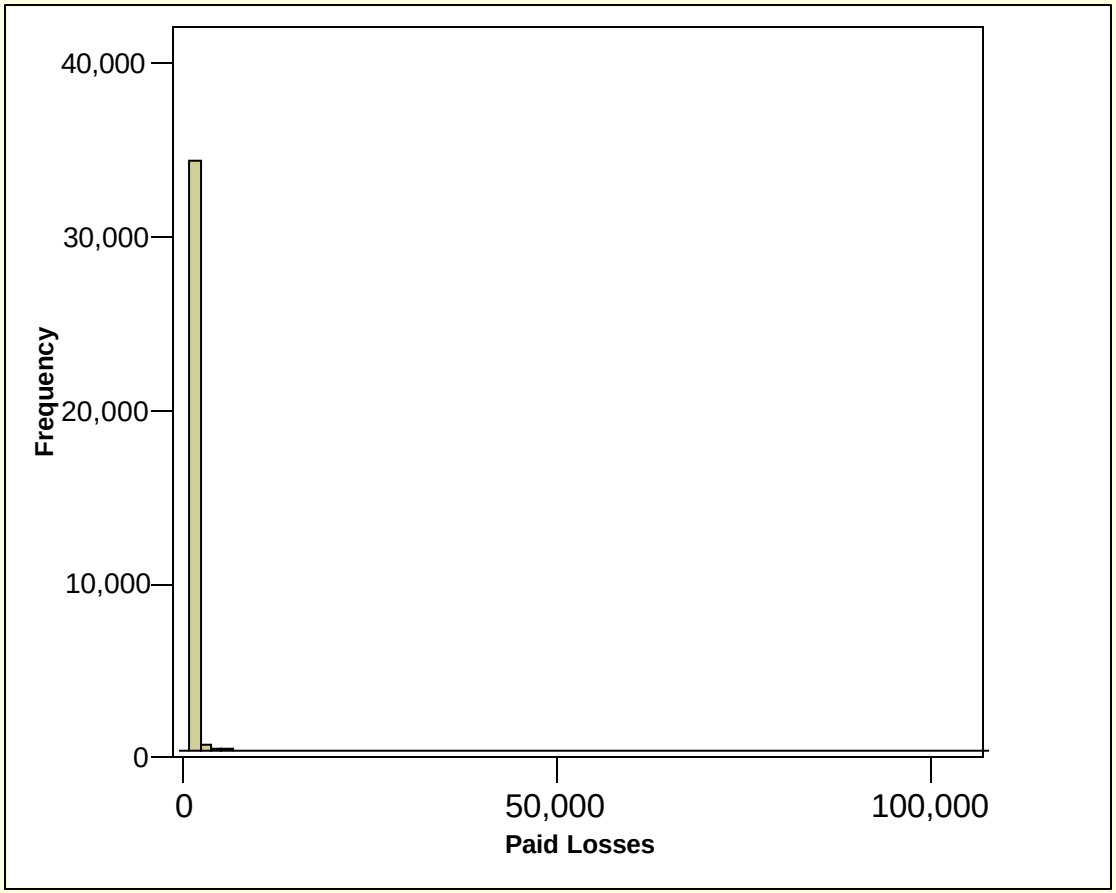
N=sample size

h =window width

Example of Suspicious Value

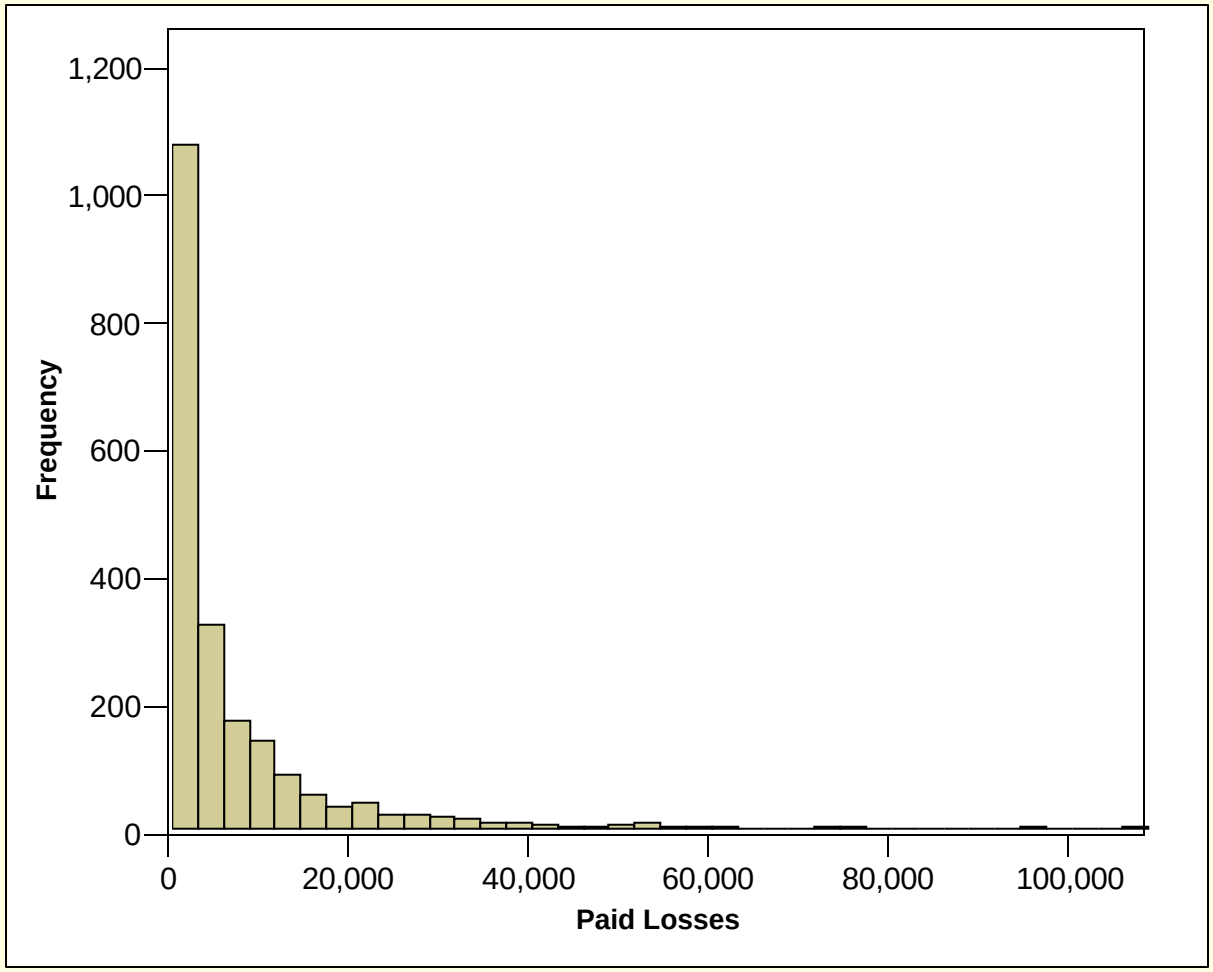


Discrete-Numeric Data



Filtered Data

Filter out Unwanted Records



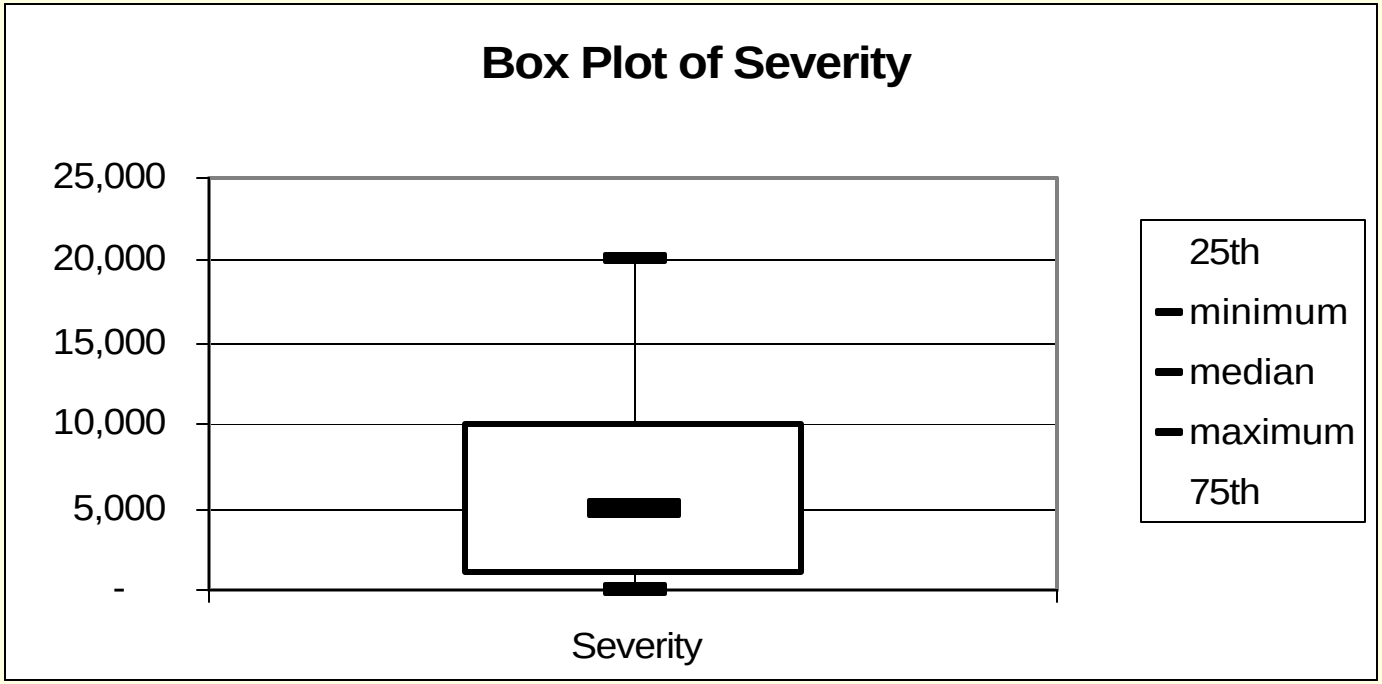
Box Plot Basics: Five – Point Summary

- Minimum
- 1st quartile
- Median
- 3rd quartile
- Maximum

Functions for five point summary

- =min(data range)
- =quartile(data range1)
- =median(data range)
- =quartile(data range,3)
- =max(data range)

Simple Box Plot

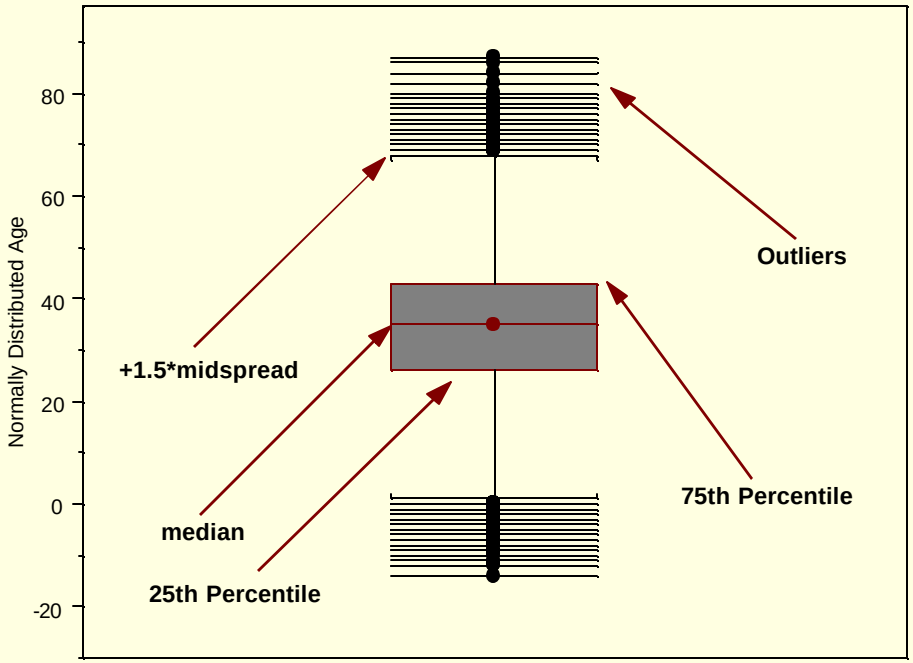


	Severity
25th	1,000
minimum	-
median	5,000
maximum	20,000
75th	10,000

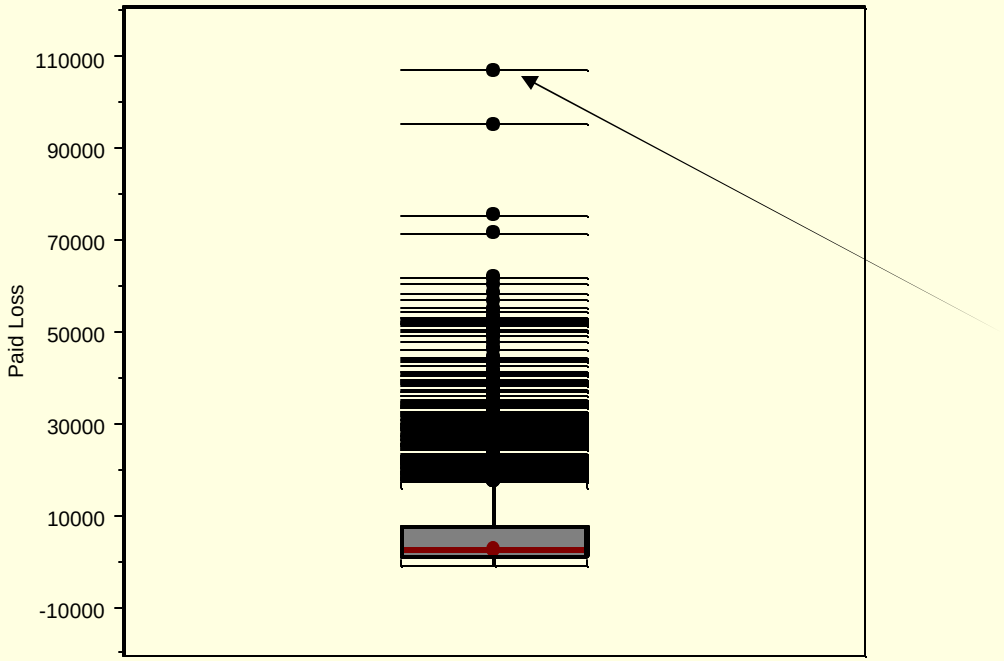
Box and Whisker Plot

- Indentifies outliers
- Need
  - Interqualtile range = 75^{th} – 25^{th} percentile, or
  - Standard deviation
- An outlier is beyond 1.5 * interquartile range, or
- Sometimes 2 * standard deviation
- Outliers represent
  - Extreme values or
  - Errors

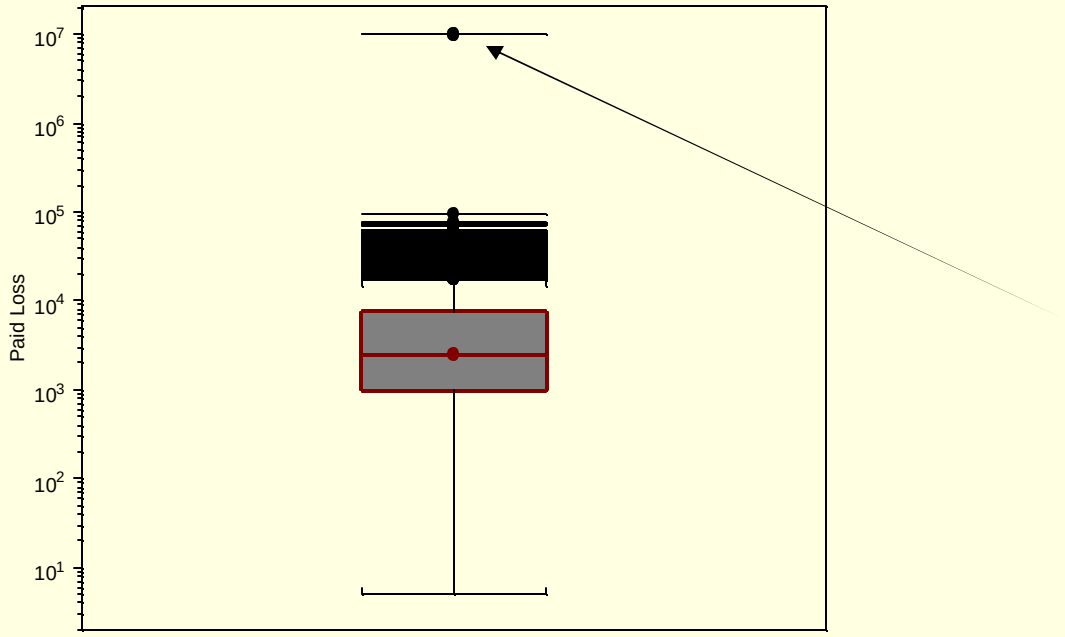
Box and Whisker Plot



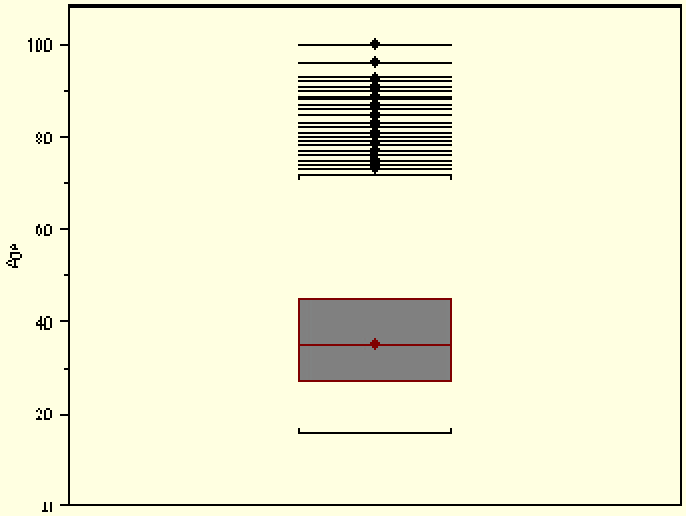
Plot of Heavy Tailed Data
Paid Losses

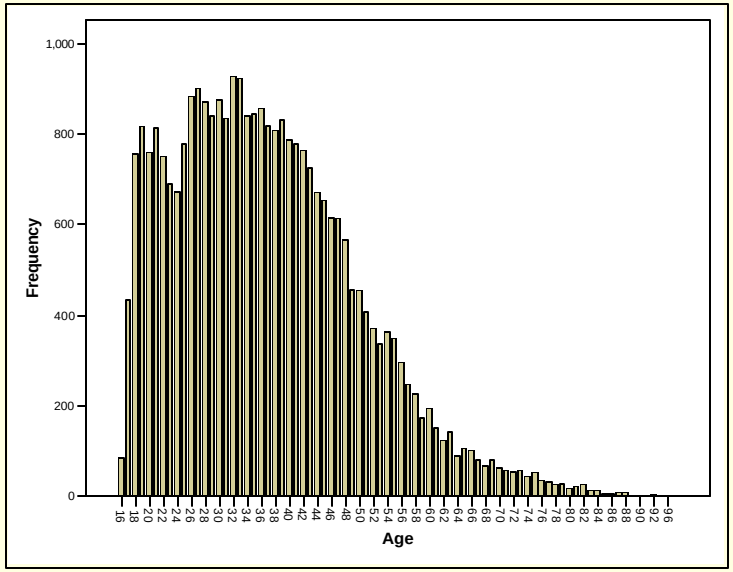


Heavy Tailed Data – Log Scale

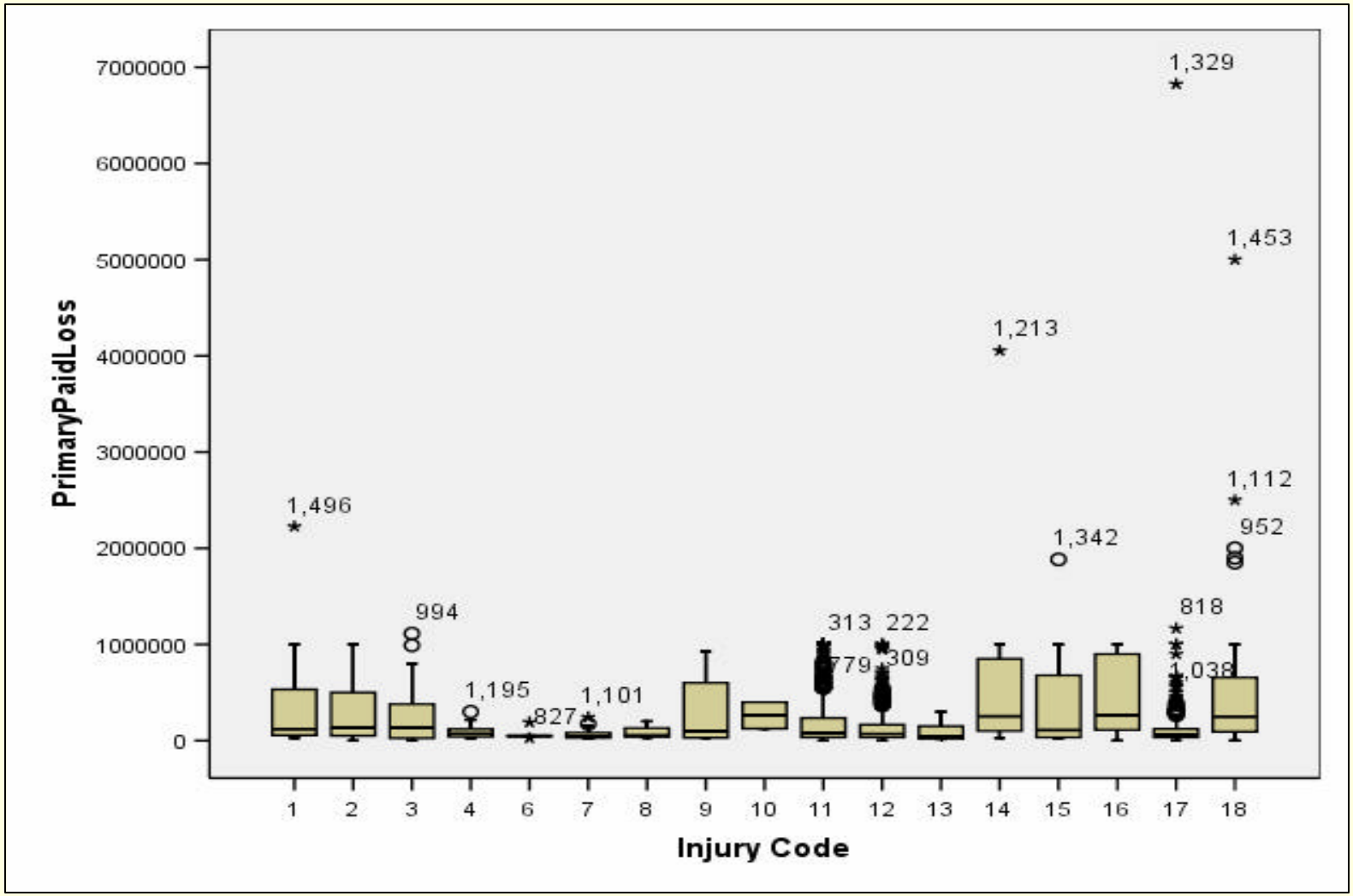


Box and Whisker Example





Multivariate Box and Whisker Plots



Descriptive Statistics

Analysis ToolPak

Statistic	*Policyholder Age*
Mean	36.9
Standard Error	0.1
Median	35.0
Mode	32.0
Standard Deviation	13.2
Sample Variance	174.4
Kurtosis	0.5
Skewness	0.7
Range	84
Minimum	16
Maximum	100
Sum	1114357
Count	30226
Largest(2)	100
Smallest(2)	16

Descriptive Statistics

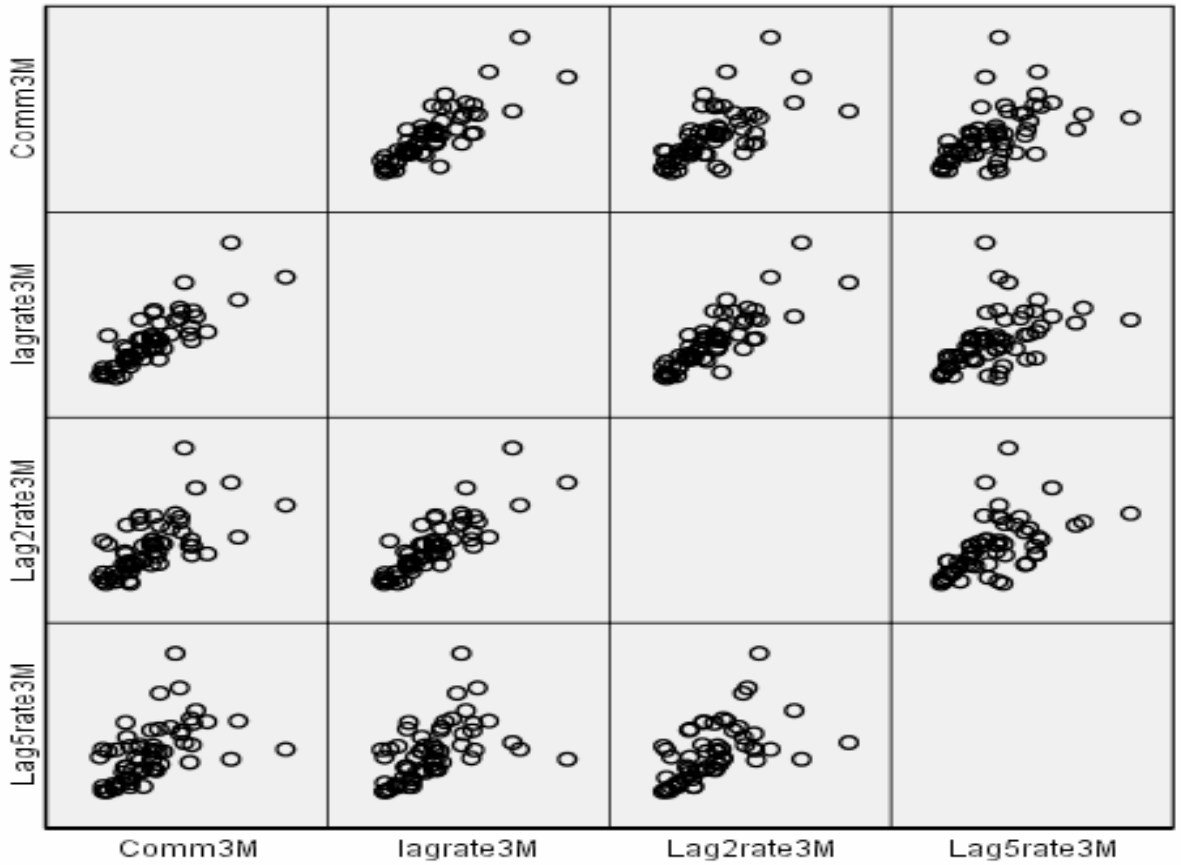
- Claimant age has minimum and maximums that are impossible

	N	Minimum	Maximum	Mean	Std. Deviation
License Year	***30,250***	***490***	***2,049***	***1,990***	***16.3***
Valid N	***30,250***				

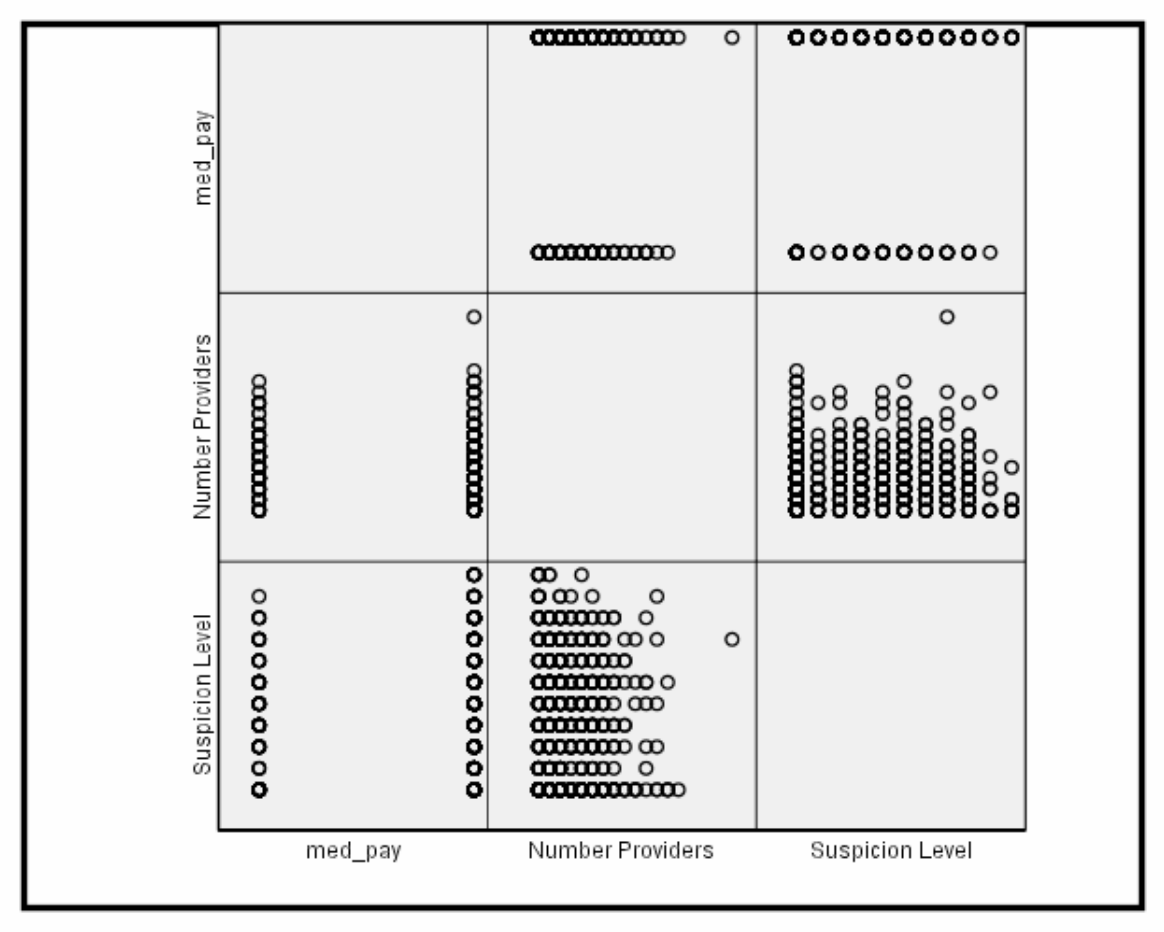
Multivariate EDA

- Often want to review relationships between multiple variables at one time
  - What structures exist?
    - What correlations exist?
  - Identify outliers

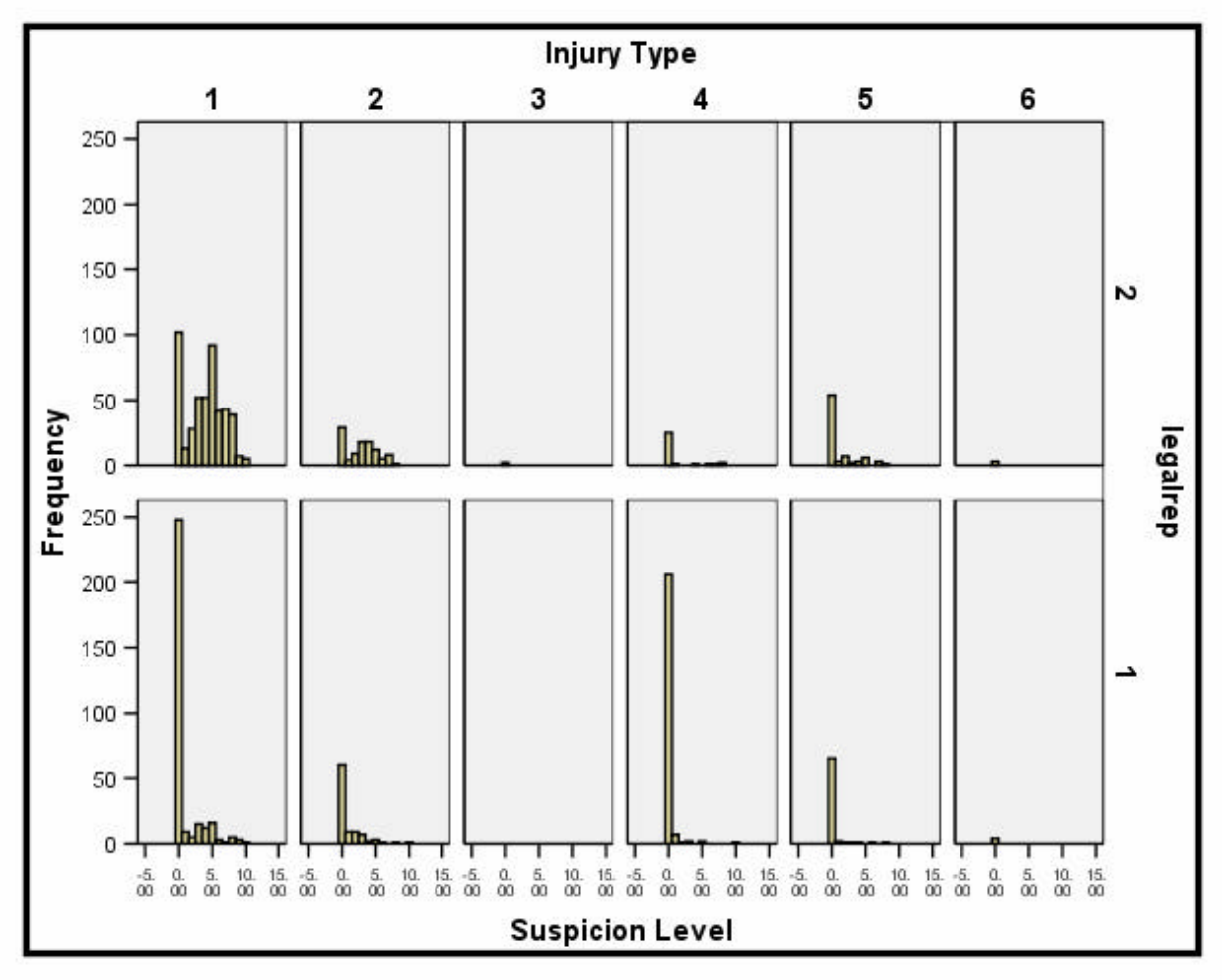
Scatterplot Matrices



Scatterplot Matrices



Panel Histogram



Data Spheres: The Mahalanobis Distance Statistic

$$\mathbf{MD} = (\mathbf{x} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

x is a vector of variables

μ is a vector of means

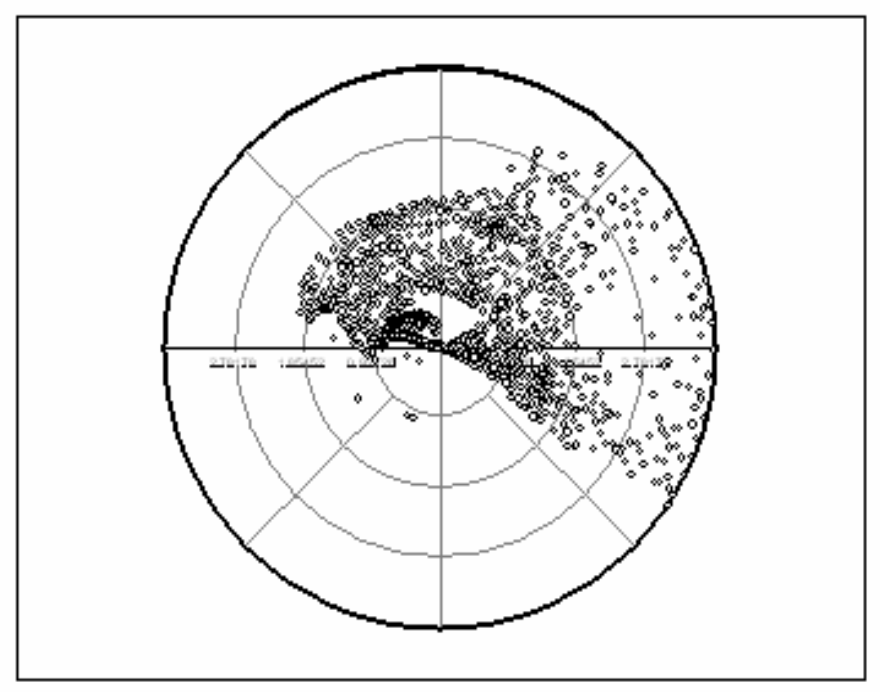
S is a variance-covariance matrix

The distance measure

1. For each variable
    1. For each record
        1. Create a standardized score by subtracting mean and dividing by standard deviation: Z=(x-mean)/sd
        2. Square it: z^2
        3. Add up the z^2 for each variable for the record: MD = $z_1^2+z_2^2\ldots+z_n^2$

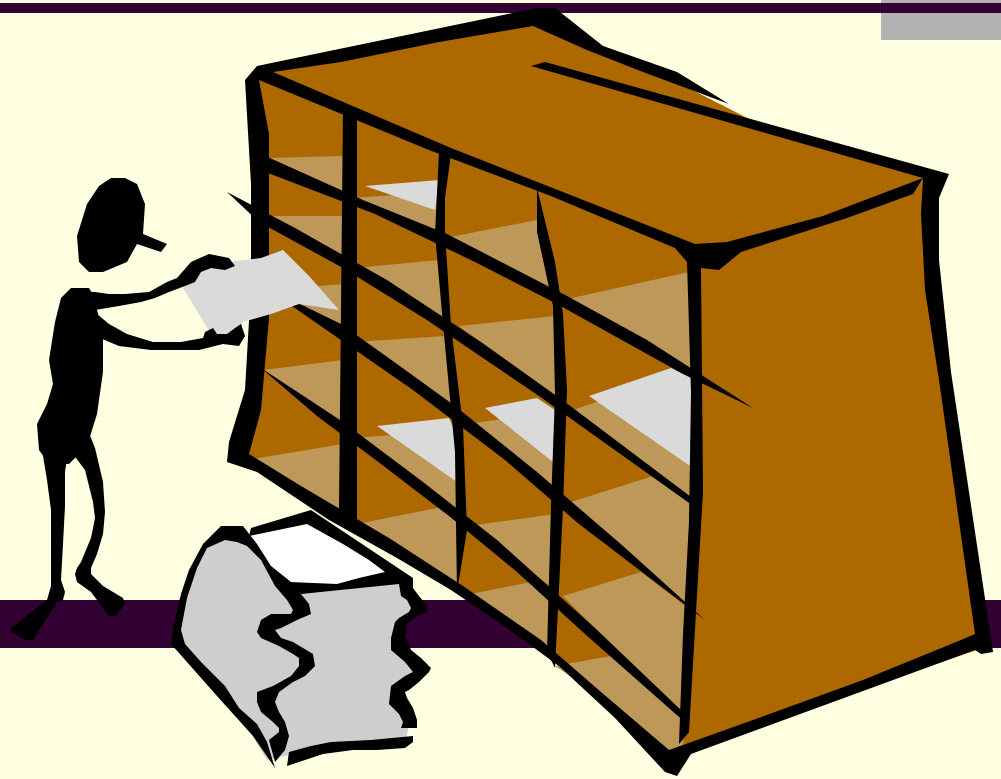
Screening Many Variables at Once

- Plot of Longitude and Latitude of zip codes in data
- Examination of outliers indicated drivers in Ca and PR even though policies only in one mid-Atlantic state



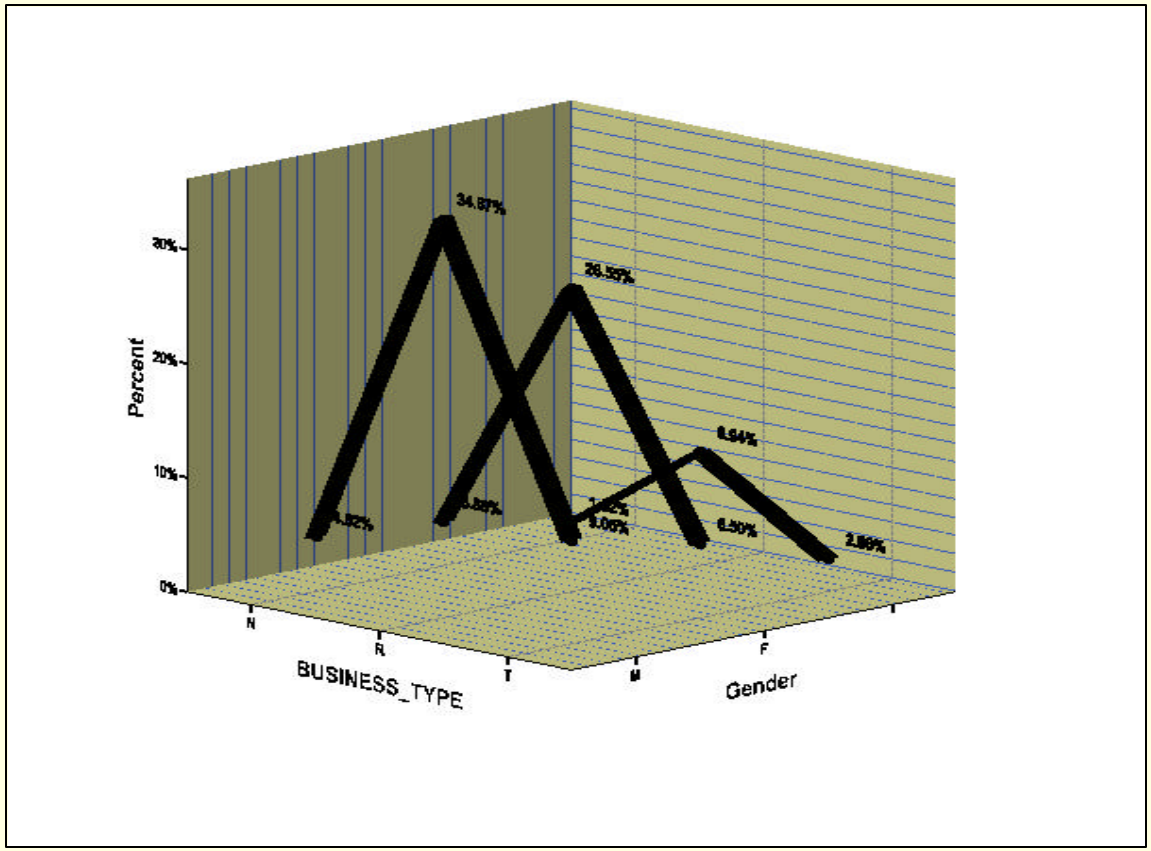
Records With Unusual Values Flagged

Policy ID	Mahalanobis Depth	Percentile of Mahalanobis	Age	License Year	Number of Cars	Number of Drivers	Model Year	Incurred Loss
22244	59	100	27	1997	3	6	1994	4,456
6159	60	100	22	2001	2	6	1993	0
22997	65	100	NA	NA	2	1	1954	0
5412	61	100	17	2003	3	6	1994	0
30577	72	100	43	1979	3	1	1952	0
28319	8,490	100	30	490	1	1	1987	0
27815	55	100	44	1976	-1	0	1959	0
16158	24	100	82	1938	1	1	1989	61,187
4908	25	100	56	1997	4	4	2003	35,697
28790	24	100	82	2039	1	1	1985	27,769



Categorical Data

Categorical Data: Data Cubes



Categorical Data

- Data Cubes
  - Usually frequency tables
  - Search for missing values coded as blanks

Gender

	Frequency	Percent
	5,054	14.3
F	13,032	36.9
M	17,198	48.7
Total	35,284	100

Categorical Data

- Table highlights inconsistent coding of marital status

Marital Status

	Frequency	Percent
	5,053	14.3
1	2,043	5.8
2	9,657	27.4
4	2	0
D	4	0
M	2,971	8.4
S	15,554	44.1
Total	35,284	100

Crosstabulations

Injury * Attorney Involvement-Plaintiff Crosstabulation

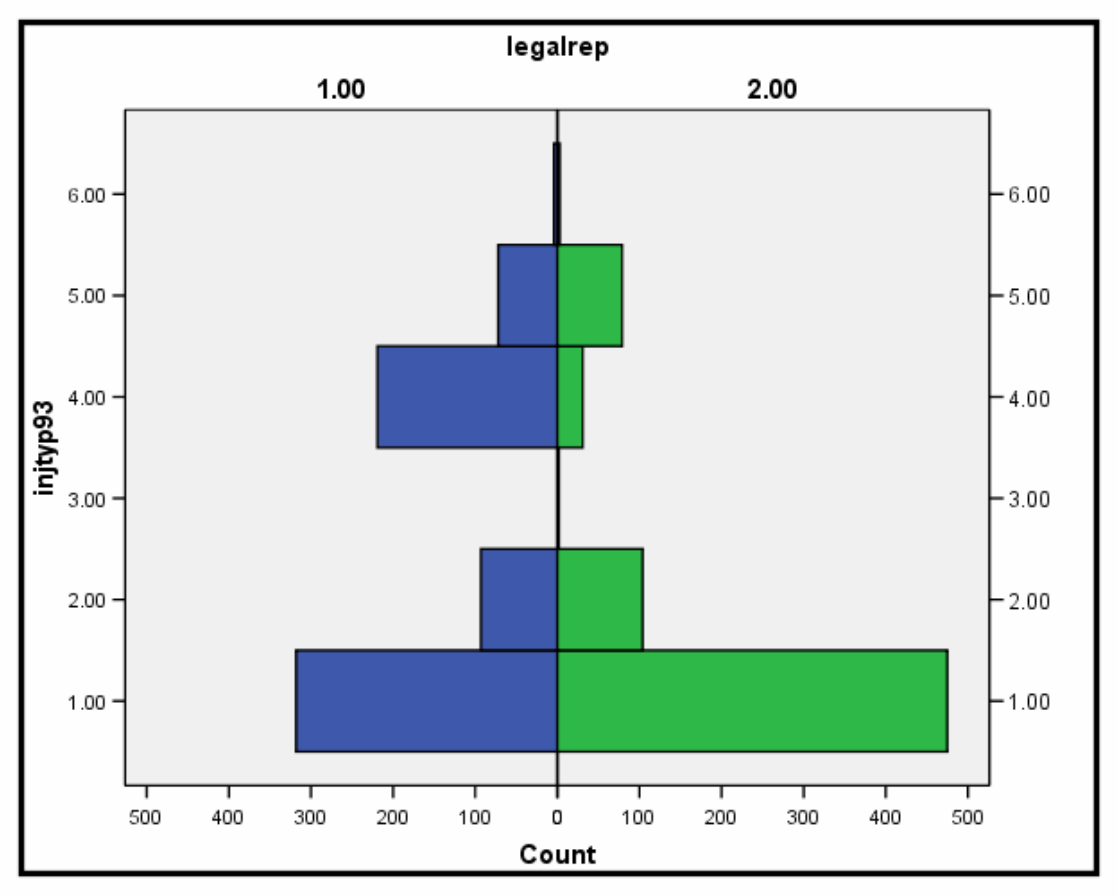
Count

		Attorney Involvement-Plaintiff		
		N	Y	Total
Injury	Amputation	1	38	39
	Backinjury	10	452	462
	Braindamage	0	26	26
	Burnschemical	0	21	21
	Burnsheat	1	44	45
	Circulatorycondition	0	2	2
	Death	4	195	199
	Eyeinjuryblindness	0	5	5
	Hearinglossorimpairment	0	6	6
	Multipleinjuries	24	597	621
	Nervouscondition	0	4	4
	Other	9	271	280
	Respiratorycondition	0	37	37
	Scarring	0	20	20
	Skindisorder	2	7	9
	Spinalcordinjuries	2	24	26
	Systemicpoisoningtoxic	0	16	16
Total		53	1765	1818

Crosstabulations

Count	AttorneyInvolvement-Insurer	
Cause	N	Y
Airtransportation		1
Drowning	1	3
Explosions	2	31
Falls	45	382
Fire	1	15
Firearm		7
Offroadvehicle	3	38
Oilgasextraction	1	42
Other_A	44	324
Othermotorvehicle	215	460
PollutionToxicexposure	35	24
Railway	3	8
Surgicalmedicalcare	3	30
Useofagriculturalmachinery		6
Useofdefectiveproduct	2	92
Grand Total	355	1463

Population Pyramid



Missing Data

Screening for Missing Data

		BUSINESS TYPE	Gender	Age	License Year
N	Valid	*35,284*	*35,284*	*30,242*	*30,250*
	Missing	*0*	*0*	*5,042*	*5,034*
Percentiles	25			*27.00*	*1,986.00*
	50			*35.00*	*1,996.00*
	75			*45.00*	*2,000.00*

Blanks as Missing

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		5,054	14.3	14.3	14.3
	F	13,032	36.9	36.9	51.3
	M	17,198	48.7	48.7	100.0
	Total	35,284	100.0	100.0	

Types of Missing Values

- Missing completely at random
- Missing at random
- Informative missing

Methods for Missing Values

- Drop record if any variable used in model is missing
- Drop variable
- Data Imputation
- Other
  - CART, MARS use surrogate variables
  - Expectation Maximization

Imputation

- A method to “fill in” missing value
- Use other variables (which have values) to predict value on missing variable
- Involves building a model for variable with missing value
  - $Y = f(x_1, x_2, \ldots x_n)$

Example: Age Variable

- About 14% of records missing values
- Imputation will be illustrated with simple regression model
  - Age = $a+b_1X_1+b_2X_2\ldots b_nX_n$

Model for Age

Tests of Between-Subjects Effects						
Dependent Variable: Age						
		Type III Sum of Squares	df	Mean Square	F	Sig.
Source	Corrected Model	3,218,216	24	134,092	1,971.2	0.000
	Intercept	9,255	1	9,255	136.0	0.000
	ClassCode	3,198,903	18	177,717	2,612.4	0.000
	CoverageType	876	3	292	4.3	0.005
	ModelYear	7,245	1	7,245	106.5	0.000
	No of Vehicles	2,365	1	2,365	34.8	0.000
	No of drivers	3,261	1	3,261	47.9	0.000
	Error	2,055,243	30,212	68		
	Total	46,377,824	30,237			
	Corrected Total	5,273,459	30,236			

Missing Values

- A problem for many traditional statistical models
  - Elimination of records missing on anything from analysis
- Many data mining procedures have techniques built in for handling missing values
- If too many records missing on a given variable, probably need to discard variable

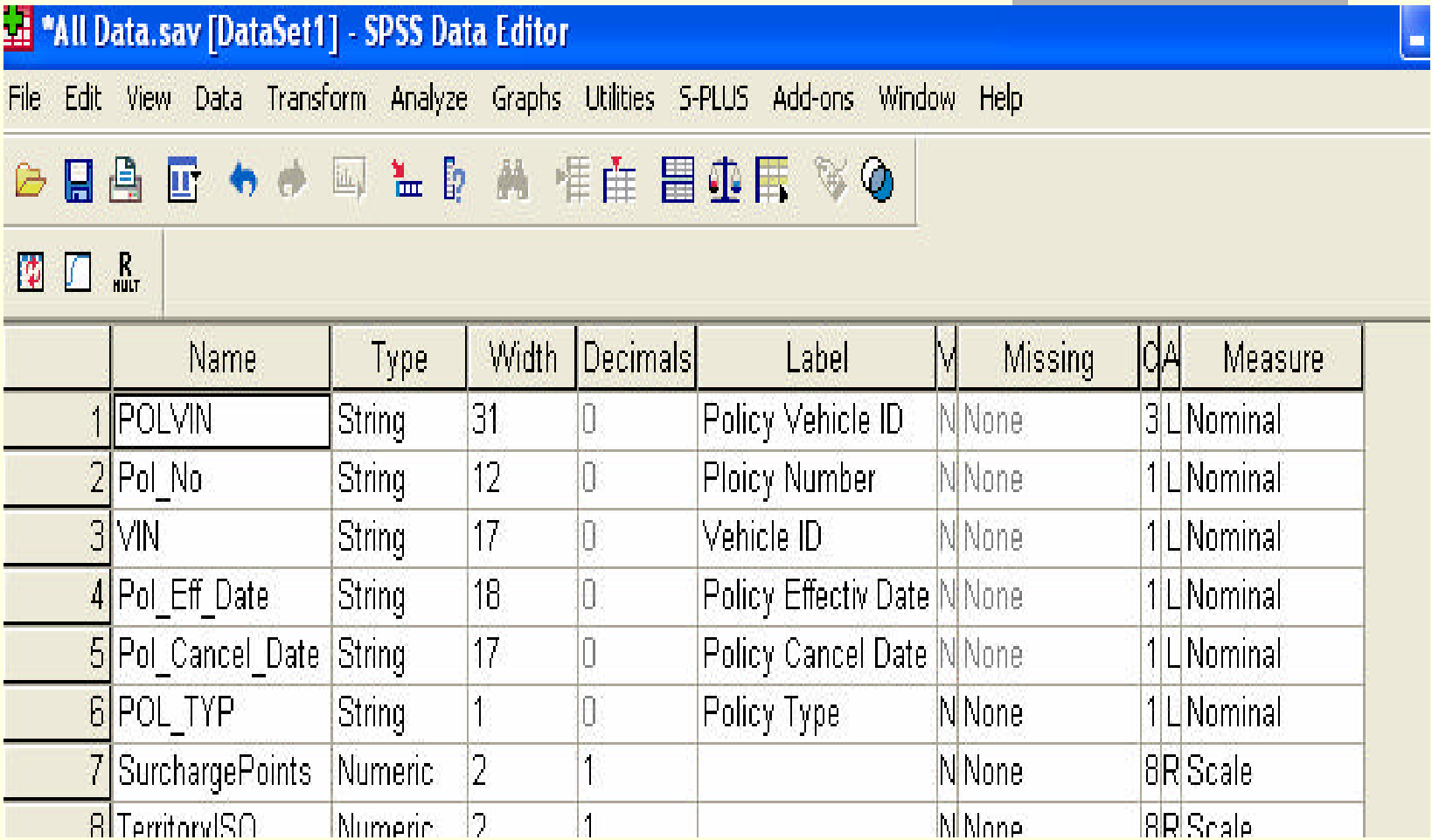
Metadata

Metadata

- Data about data
  - A reference that can be used in future modeling projects
- Detailed description of the variables in the file, their meaning and permissible values

Marital Status Value	Description
1	Married, data from source 1
2	Single, data from source 1
4	Divorced, data from source 1
D	Divorced, data from source 2
M	Married, data from source 2
S	Single, data from source 2
Blank	Marital status is missing

Metadata: Sources of Documentation of Data



	Name	Type	Width	Decimals	Label	V	Missing	C	A	Measure
1	POLVIN	String	31	0	Policy Vehicle ID	N	None	3	L	Nominal
2	Pol_No	String	12	0	Ploicy Number	N	None	1	L	Nominal
3	VIN	String	17	0	Vehicle ID	N	None	1	L	Nominal
4	Pol_Eff_Date	String	18	0	Policy Effectiv Date	N	None	1	L	Nominal
5	Pol_Cancel_Date	String	17	0	Policy Cancel Date	N	None	1	L	Nominal
6	POL_TYP	String	1	0	Policy Type	N	None	1	L	Nominal
7	SurchargePoints	Numeric	2	1		N	None	8	R	Scale
8	TerritoryISO	Numeric	2	1		N	None	8	R	Scale

Tx WC Dat2 Rev : Table

Field Name	Data Type
EXTSEQ	Number
TYPEF	Text
AccDate	Date/Time
RptDate	Date/Time
SuitDate	Date/Time
TrialDate	Date/Time
SEtlDate	Date/Time
CloseDate	Date/Time
Initialindemnityreserve	Number
Initialexpensereserve	Number
InitialexpenditurereserveQ8A	Number
Finalindemnityreserve	Number
Finalexpensereserve	Number
FinalexpenditurereserveQ8D	Number
Attorneyinvolvementplaintiff	Text
Attorneyinvolvementinsurer	Text
Attorneyinvolvementinsured	Text
Legalstagewheresettlementw	Number
Totalcourtverdictamount	Number
Amountofsettlement	Number
PrimaryPaidLoss	Number

General | Lookup

Field Size	Double
Format	General Number
Decimal Places	Auto
Input Mask	
Caption	
Default Value	
Validation Rule	
Validation Text	
Required	No
Indexed	No

Partitioning Data

Partition Data

- Training Sample
- Testing Sample

or

- Training Sample
- Validation sample
- Testing sample

Random Selection of Data

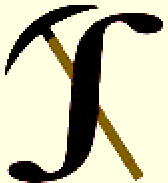
Random Number
0.507475872
0.569931387
0.52819047
0.076093765
0.776281923
0.258076057

Sampling Approaches

- Random
  - Every record has same probability of occurring
- Stratified
  - Some records (i.e., target variable that occurs infrequently in data) have higher probability of occurring

Sampling for Data Reduction

- Data set is too large to perform analysis with
- This may be dependent on software you are working with
- Must be careful not to under-sample values that occur infrequently by are important to analysis



Normalization and Scaling

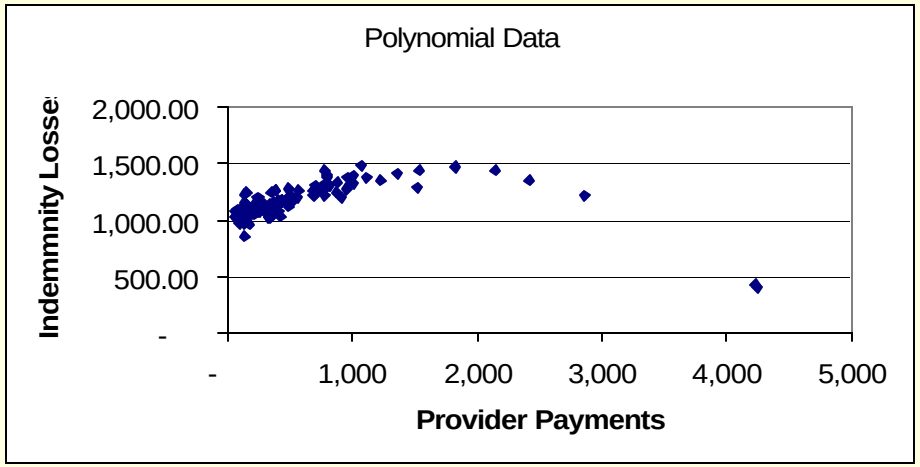
Two Common Ways to Standardize variables

- Create z-score
  - z=(x-mean)/sd
  - Common in factor analysis

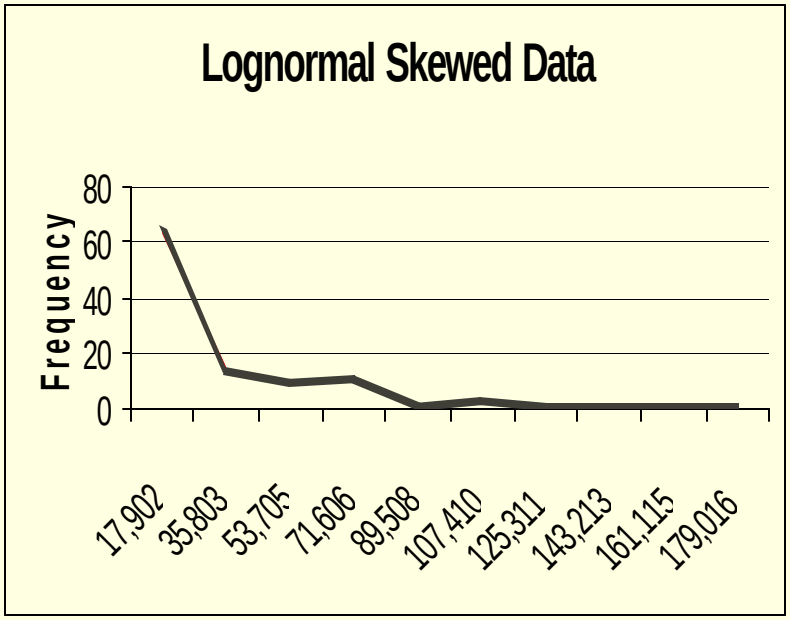
- Create a number that varies between zero and one or between minus one and one
  - Common in neural networks
  - Compute range (max-min)
  - Subtract minimum and divide by range
  - Scaled var=(x-min)/range

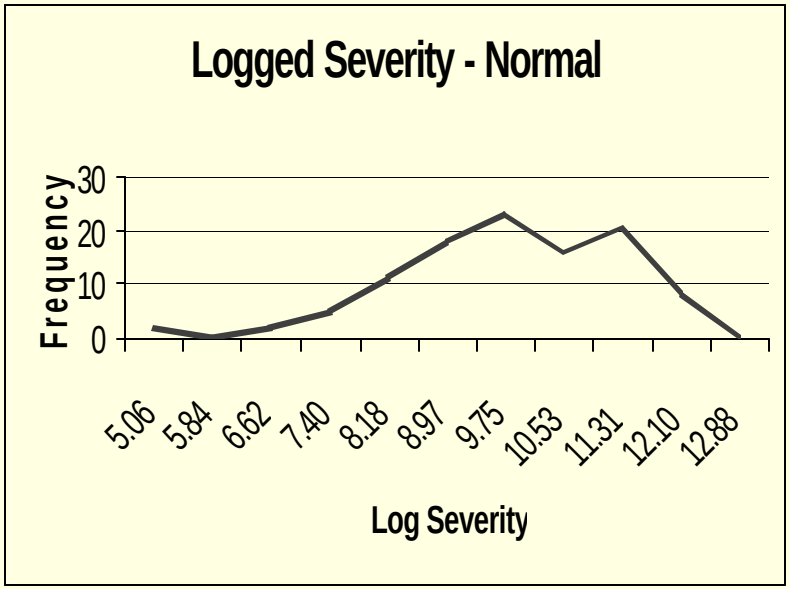
Transformations

- Most common one: Logarithms
  - X'=log(x), x>0
- Also polynomial transformation
  - Create X^2, X^3 etc.



Original and Logged Data for Skewed Distribution



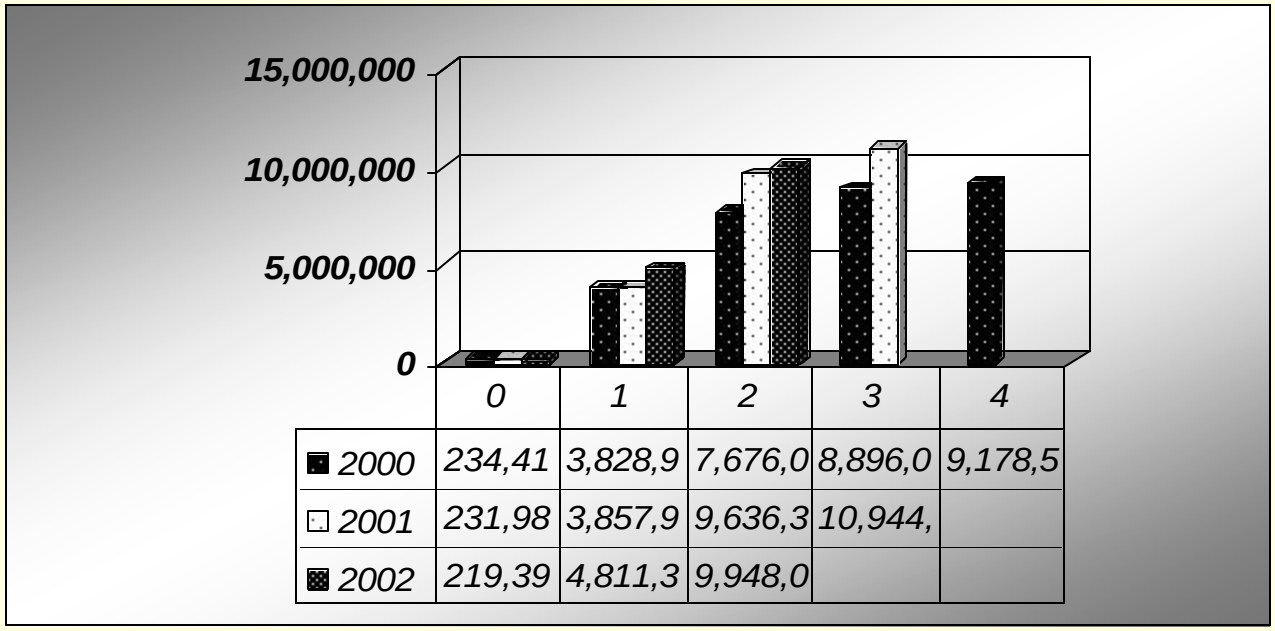


Why Normalize Variables?

- Everything is on the same scale
- A variable does not dominate model because its units are higher

Other Data Adjustments

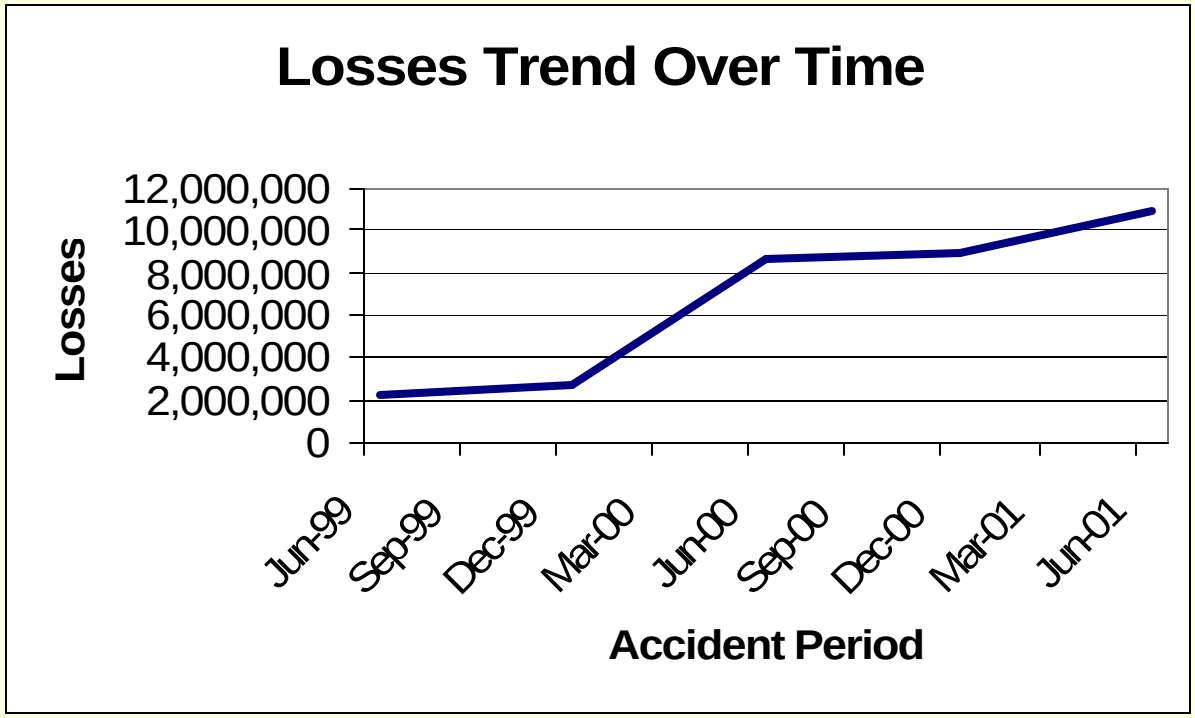
- Loss Development



	0	1	2	3	4
2000	234,41	3,828,9	7,676,0	8,896,0	9,178,5
2001	231,98	3,857,9	9,636,3	10,944,	
2002	219,39	4,811,3	9,948,0		

Other Data Adjustments

- Trend



Trends

- Frequency
- Severity
- Benefit level

Categorical Variable Mapping

- Binary Dummy Variables

Injury	Injury Dummy 1	Injury Dummy 2	Injury Dummy 3	Injury Dummy 4	Injury Dummy 5
Amputation	1	0	0	0	0
Amputation	1	0	0	0	0
Backinjury	0	1	0	0	0
Multipleinjuries	0	0	1	0	0
Death	0	0	0	1	0
Backinjury	0	0	0	0	1
Amputation					

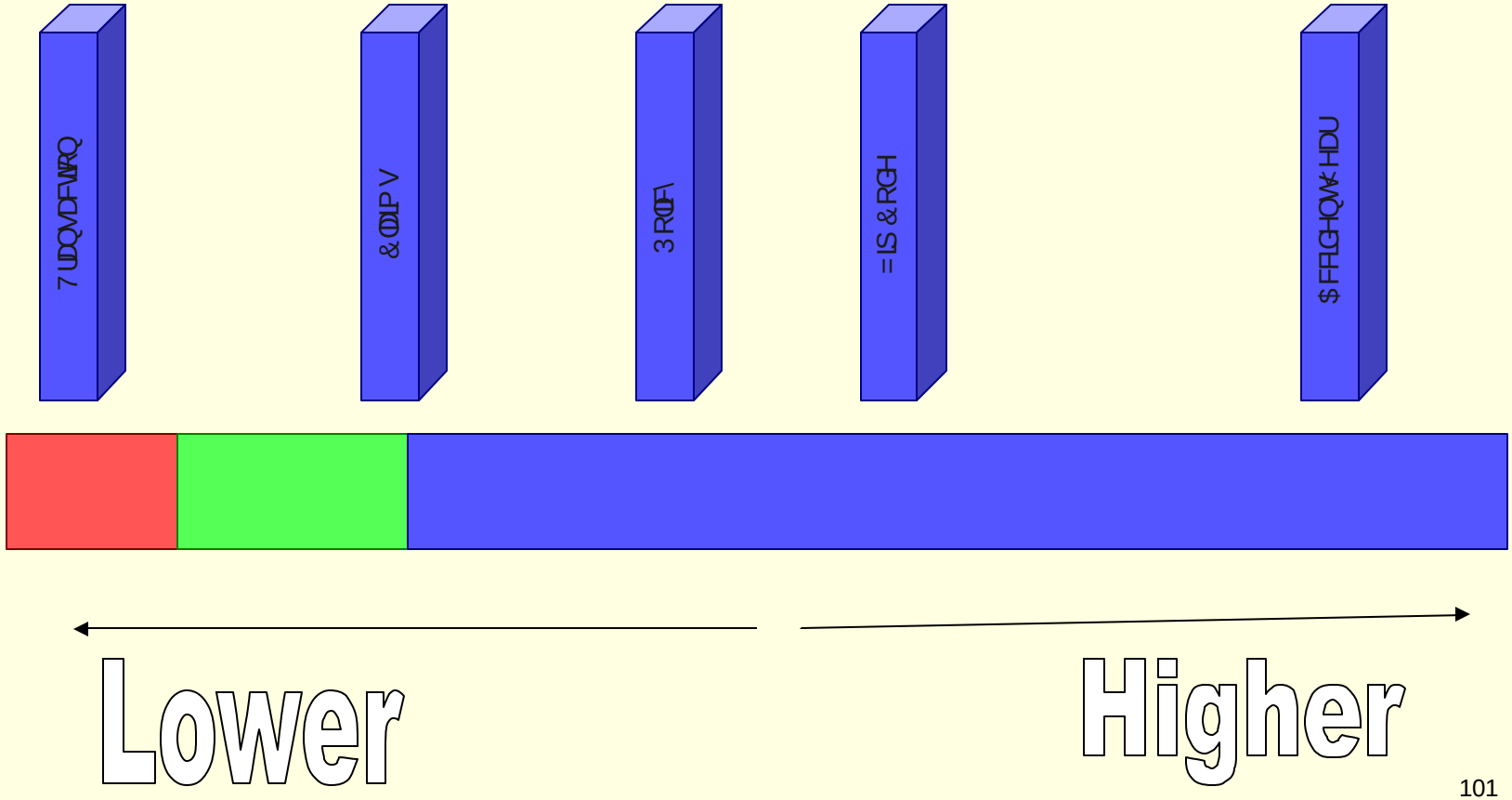
Categorical Variable Mapping

- Use domain knowledge to map categories into new variables
- Example
  - Map state to region, such as northeast, midwest
  - Map injuries to high severity, moderate severity, low severity

Density and Sparsity

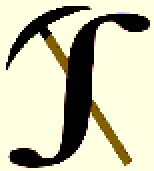
Count	Serious		
Cause of Loss	0	1	Grand Total
Airtransportation	1		1
Drowning	4		4
Explosions	16	17	33
Falls	351	76	427
Fire	9	7	16
Firearm	5	2	7
Offroadvehicle	28	13	41
Oilgasextraction	27	16	43
Other_A	276	92	368
Othermotorvehicle	554	121	675
PollutionToxicexposure	58	1	59
Railway	8	3	11
Surgicalmedicalcare	27	6	33
Useofagriculturalmachinery	6		6
Useofdefectiveproduct	71	23	94
Grand Total	1441	377	1818

Aggregation Level



Variables Derived from Aggregation

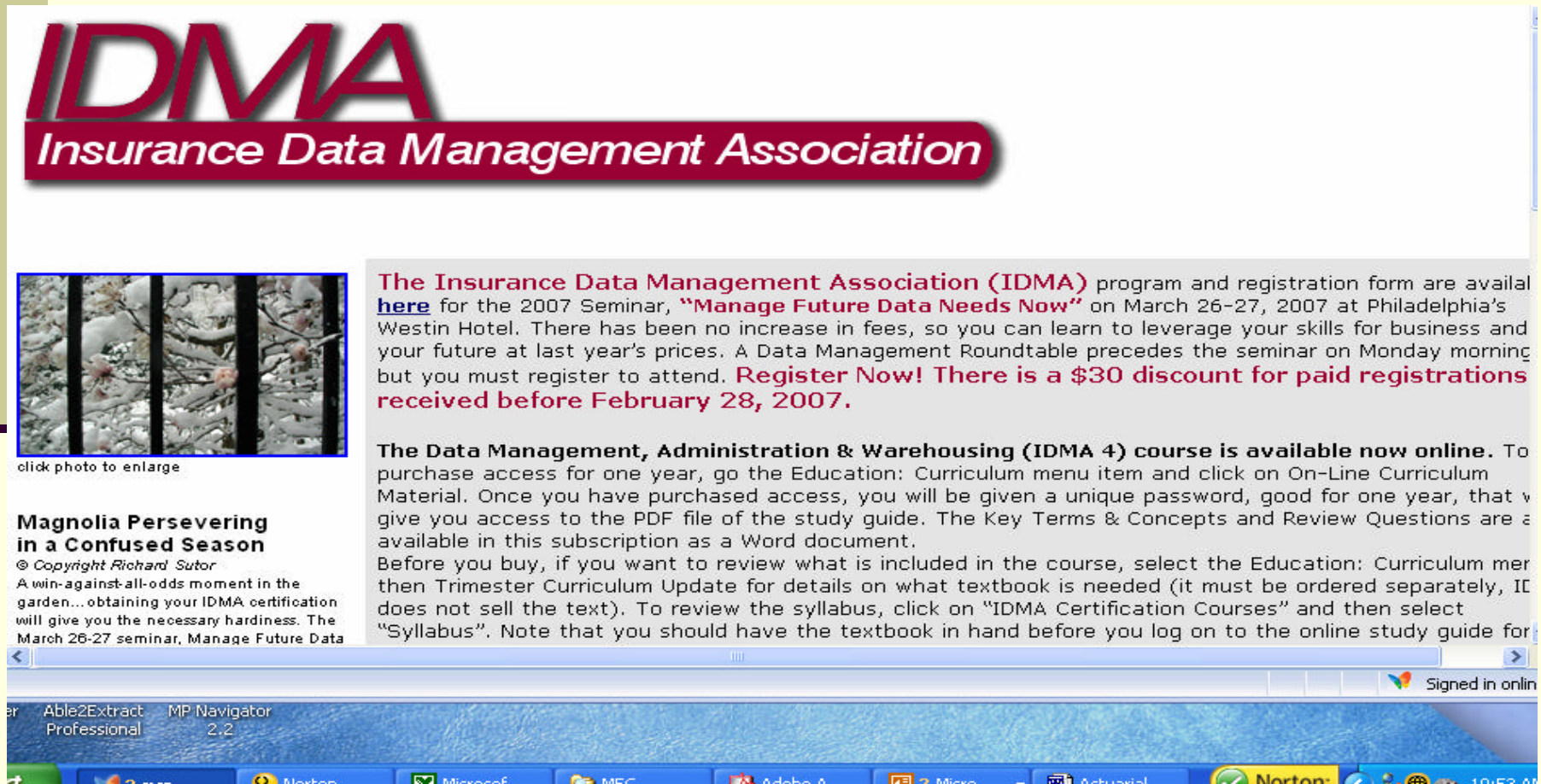
- Variables can be created from internal data through aggregation
  - Number of injuries by class code
  - Number of litigated claims per zip code
  - Severity of loss by injury code and state
- Additional derived variables can be created using external data



Resources for Data Preparation

Resources: Data Quality

- IDMA web site: www.idma.org



IDMA: Data Quality Resource

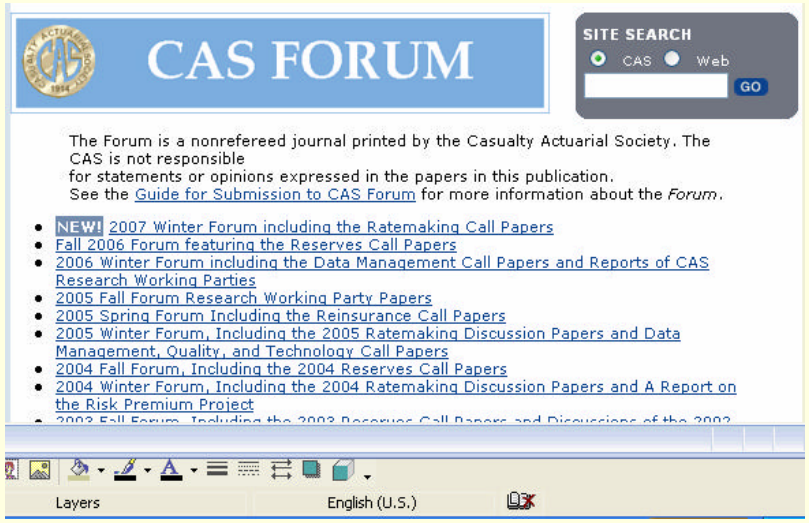
Value Proposition

With every advance in technology, the value of data increases. Evolving and new technologies make sharir data fairly simple, but if bad information moves through a process, the costs grow exponentially at each st Data must be managed to ensure quality, to ensure maximum benefit to the organization. If Senior Manage receive bad data, how can they properly manage? If Marketing receives incorrect information, the dollar va of lost opportunities can only be guessed. In the realm of regulatory or statutory reporting, bad data may create a liability for your organization. All of the Actuarial, Claims, Compliance, Finance, and Information Technology functions need reliable data to work with. The data manager's function is to provide other man with the information they need to fulfill their role.

General Information	Value to Actuaries	Value to Marketing
Value to Senior Management	Value to Claims	Value to Statistical and Regulatory Reporting
Value to Compliance/ Government	Value to Finance	Value to Underwriting
Relations and Regulators	Value to Information Technology	

Casualty Actuarial Society Data Quality Educational Materials Working Party

- Go to www.casact.org, type in Data Quality
- CAS Winter Forum:http://www.casact.org/pubs/forum/
- http://www.casact.org/pubs/forum/07wforum/07w279.pdf for Survey od Data Quality Texts



Survey of Data Management and Data Quality Texts

Author	Section	Data Quality	Principles of Data Quality	Metadata	Exploratory Data Analysis	Data Audits
Olsen	3.1	●●●●○	●●●●○	●●●○○	●●○○○	●●●○○
Dasu	3.2	●●●●○	●●●○○	●●●○○	●●●●○	●○○○○
English	3.3	●●○○○	●●●◖○	●●●◖○		
Loshin	3.4	●●●○○	●●●○○	●●●○○	●●●○○	●●○○○
Inmon	3.5	●○○○○	●●○○○	●●●○○		
Redman	3.6	●●●○○	●●○○○		●●●○○	
Watson	3.7	●●●●○	●●●●○	○○○○○		
Kit	3.8	○○○○○			○○○○○	○○○○○
IDMA	3.9	○○○○○	○○○○○		○○○○○	○○○○○

Draft Paper on Information Quality

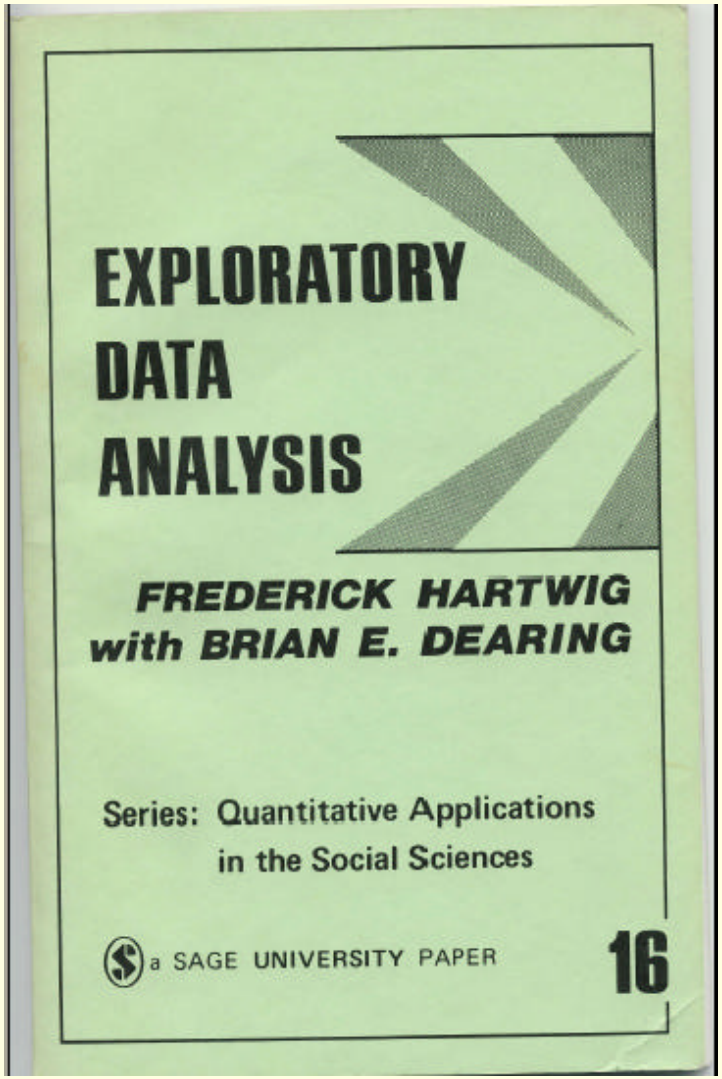
Actuarial I.Q. (Information Quality)

CAS Data Management Educational Materials Working Party

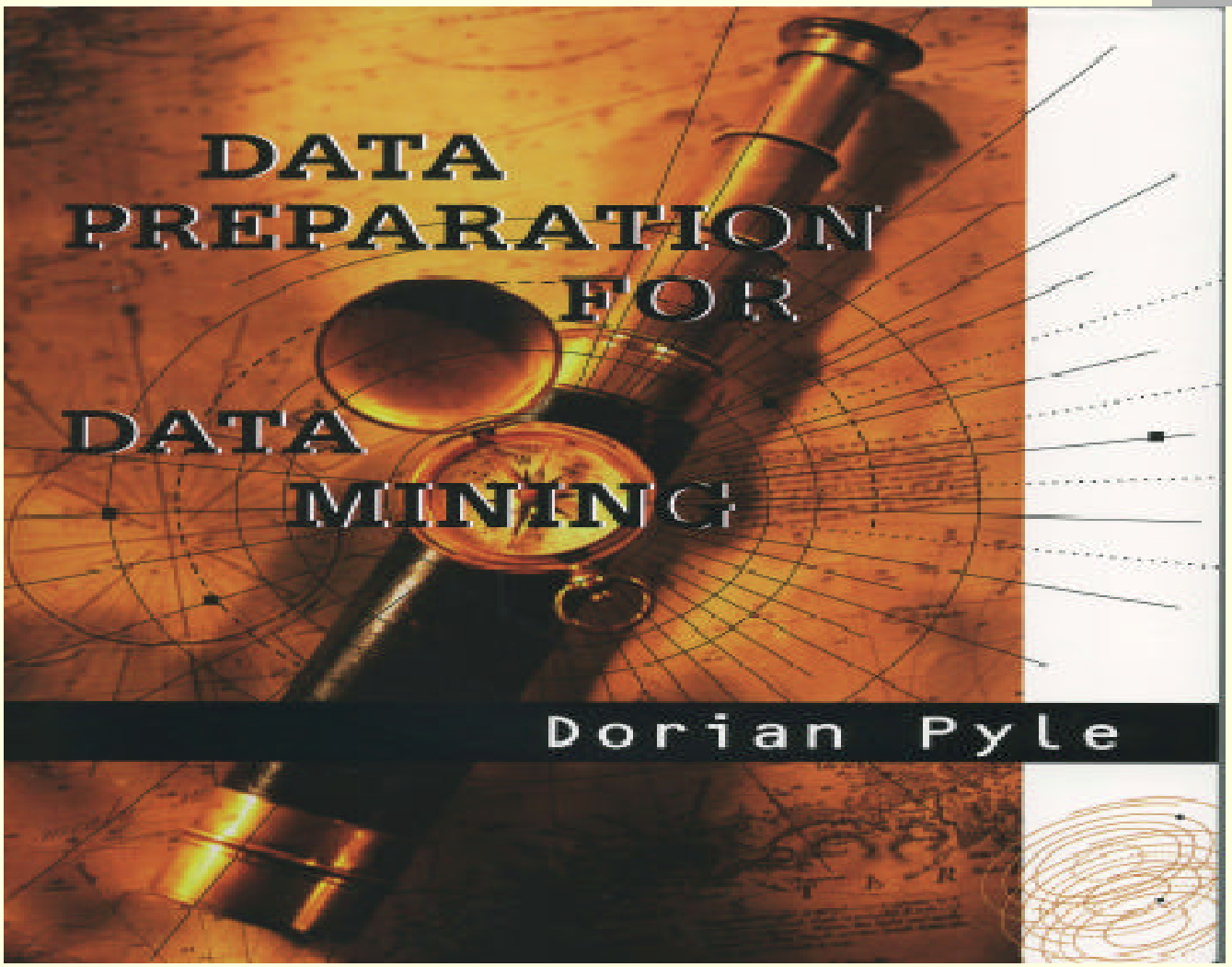
1. INTRODUCTION

Data quality is a significant concern for most actuaries. In Britain, a GIRO Data Quality working party survey [1] found that about 25% of actuaries' time is expended on data quality issues. The survey also found that about 30% of actuarial analyses are adversely affected by data quality problems. However actuaries, as both key consumers and providers of information, are uniquely well-positioned to deal with the pervasiveness of poor data quality in insurance.

Exploratory Data Analysis



Data Preparation



Data Preparation Words of Wisdom from Paul Pries of General Casualty

Work Comp data considerations

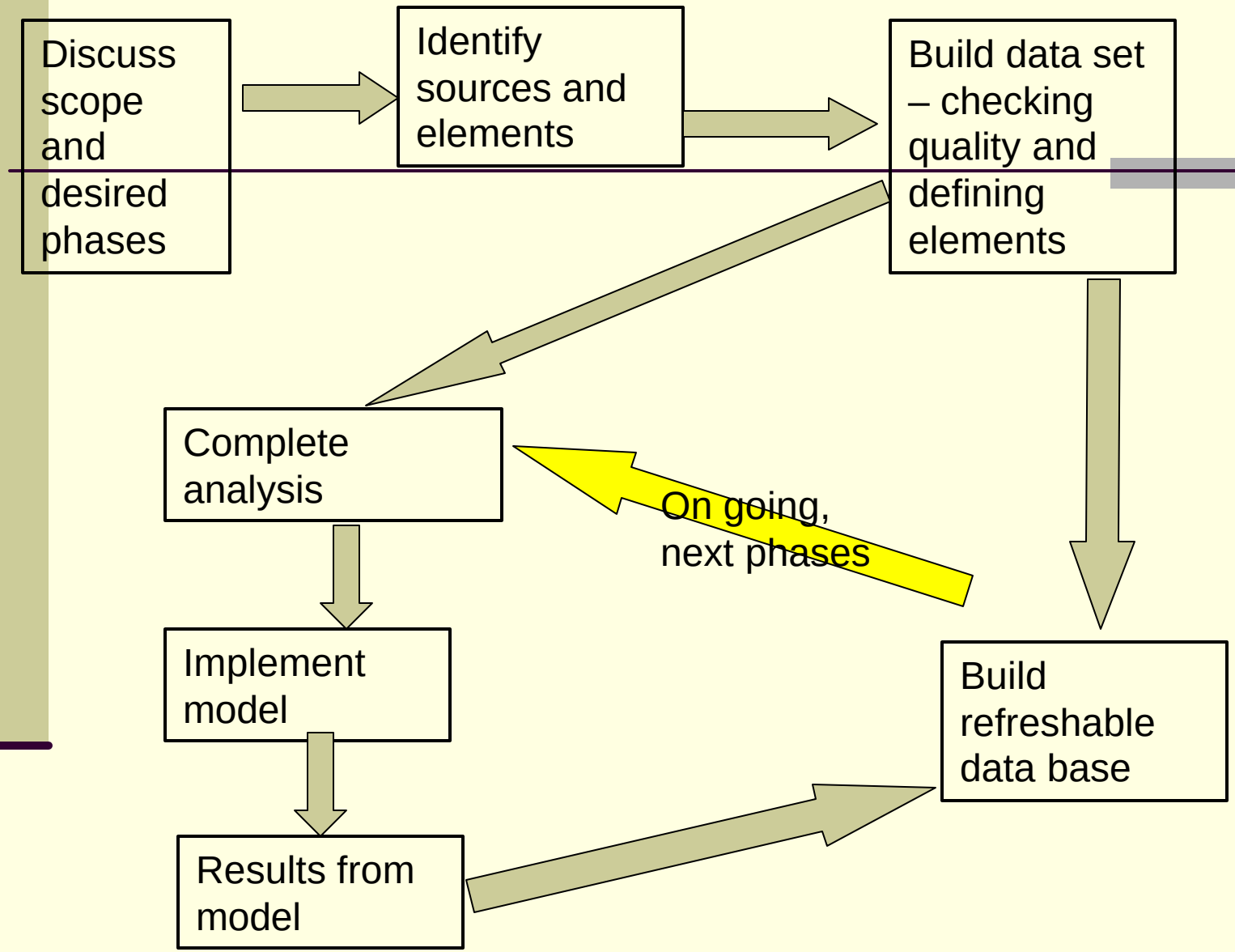
- What is the project trying to accomplish?
- Develop an overall game plan
  - Data gathering
  - Analysis
- Recommendation – if you plan to do multiple lines – make them separate projects.
- Understand the tools that will be used for analysis.

Work Comp data considerations

- Profile your data and investigate data quality.
- Frequently recheck your data – you may start with policy data and merge on data from other sources (claims data, customer data, outside data etc.) Check data after each step.

Work Comp data considerations

- Document, document, document
  - Meta data
  - Decisions reached during analysis


Discuss scope and desired phases
Identify sources and elements
Build data set – checking quality and defining elements
Complete analysis
On going, next phases
Implement model
Build refreshable data base
Results from model

Additional Library for Getting Started

- Dasu and Johnson, *Exploratory Data Mining and Data Cleaning*, Wiley, 2003
- Francis, L.A., “Dancing with Dirty Data: Methods for Exploring and Cleaning Data”, CAS Winter Forum, March 2005, www.casact.org
- Find a comprehensive book for doing analysis in Excel such as: Jospeh Schmuller, Statistical Analysis With Excel for Dummies
- Check www,data-mines.com for related materials and data