Data Preparation
Part 1: Exploratory Data Analysis & Data Cleaning, Missing Data

CAS Predictive Modeling Seminar

Louise Francis
Francis Analytics and Actuarial Data Mining, Inc.
www.data-mines.com
Louise.francis@data-mines.cm

LF1

Objectives

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

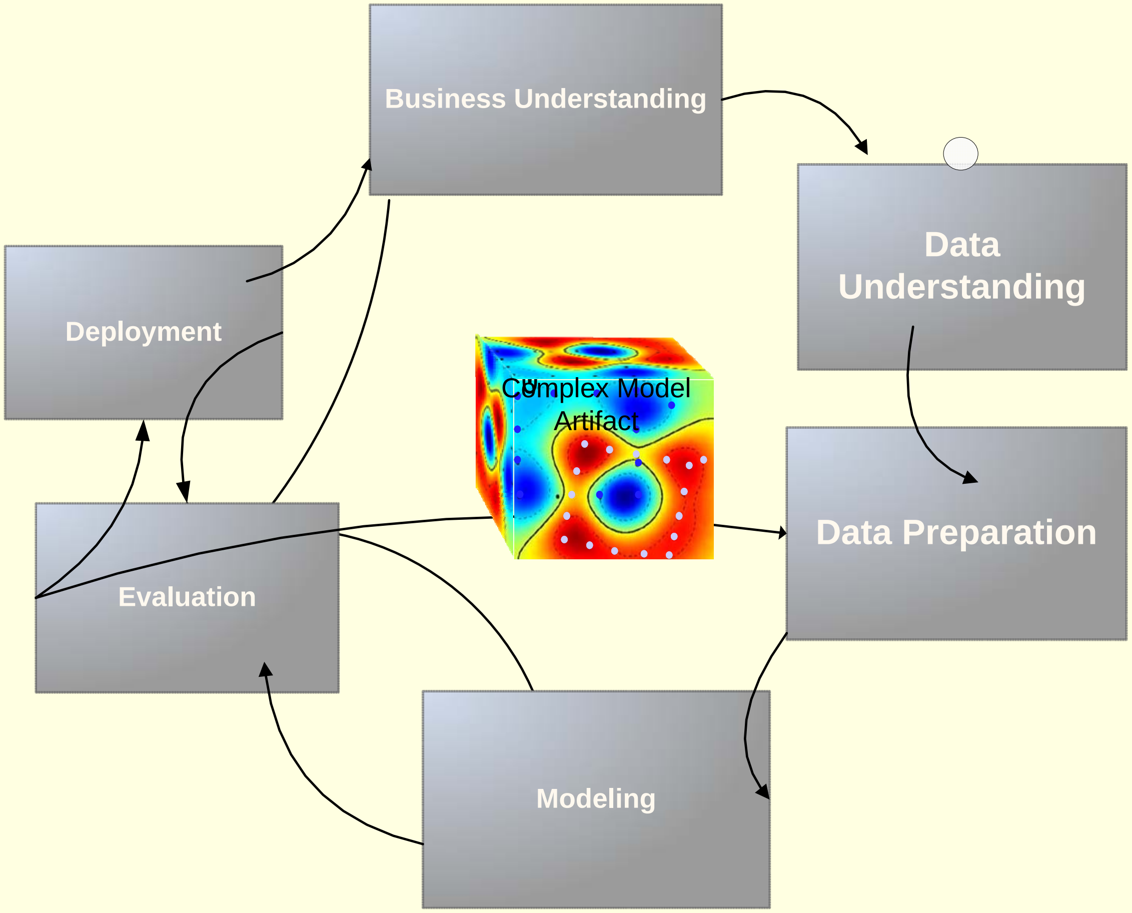
Slide 2

LF1	Louise Francis, 9/29/2006

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

Modelling Process



Business Understanding

Data Understanding

Data Preparation

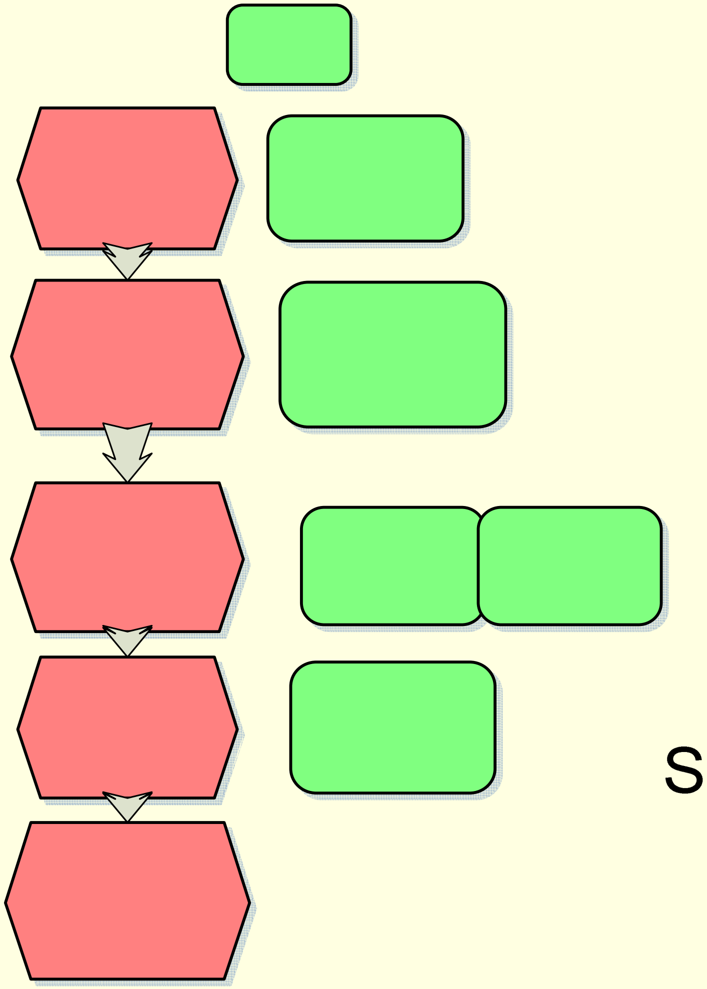
Modeling

Evaluation

Deployment

Complex Model Artifact

Data Preprocessing



Data sets

Select Data

Rational for inclusion

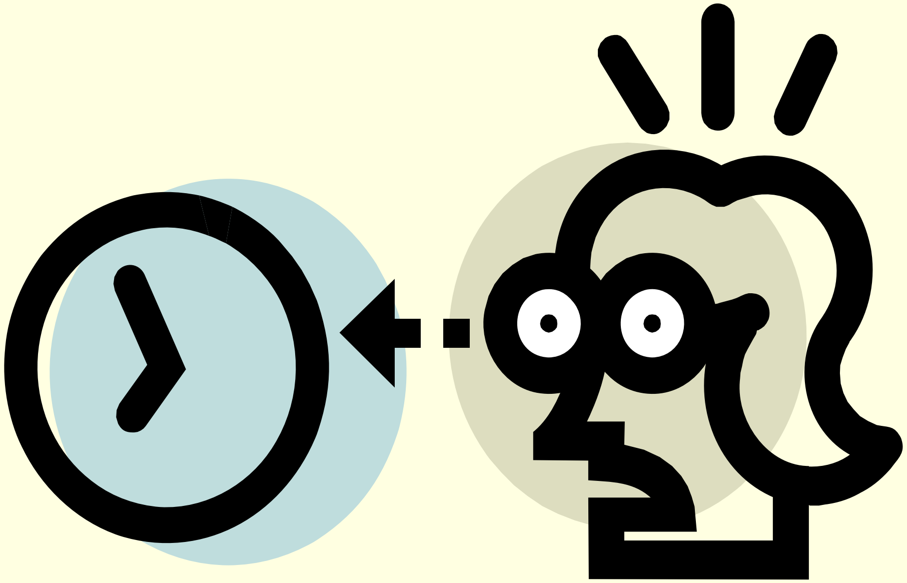
Clean Data

Data Cleaning Repor

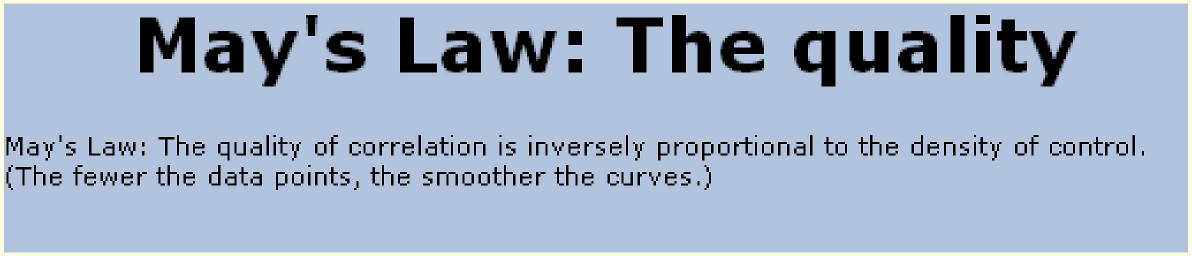
Data Quality Problem

Data Quality: A Problem

- Actuary reviewing a database



May’s Law



May's Law: The quality

May's Law: The quality of correlation is inversely proportional to the density of control.
(The fewer the data points, the smoother the curves.)

It’s Not Just Us

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

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 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

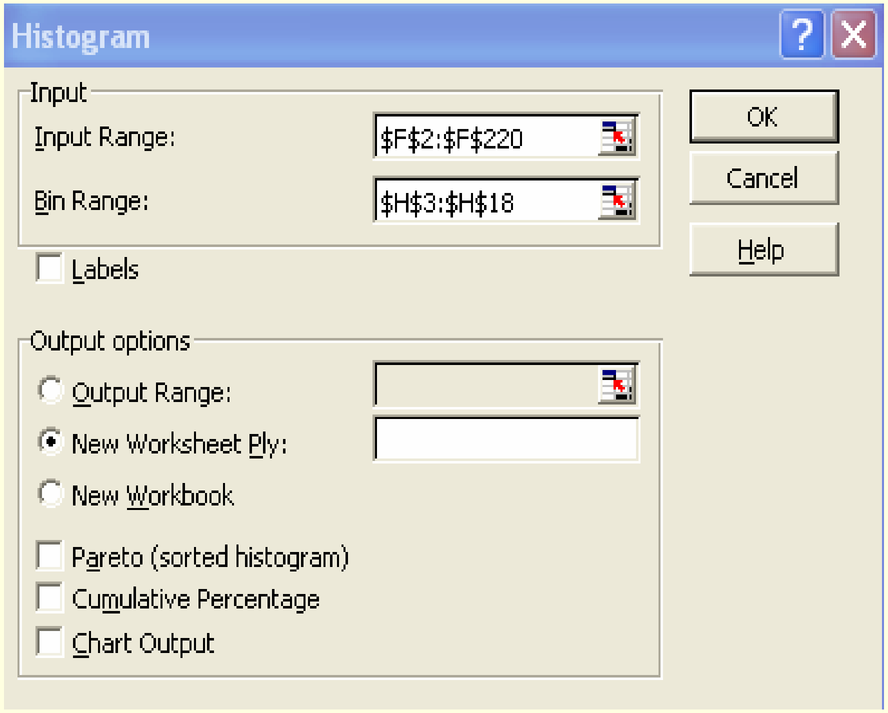
- Private passenger auto
- Some variables are:
  - Age
  - Gender
  - Marital status
  - Zip code
  - Earned premium
  - Number of claims
  - Incurred losses
  - Paid losses

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



Input Range: $F$2:$F$220

Bin Range: $H$3:$H$18

□ Labels

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

□ Chart Output

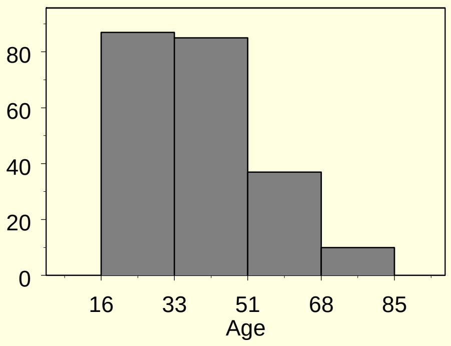
Histograms

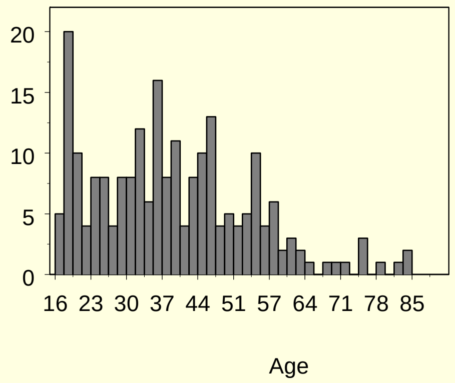
Frequencies for Age Variable

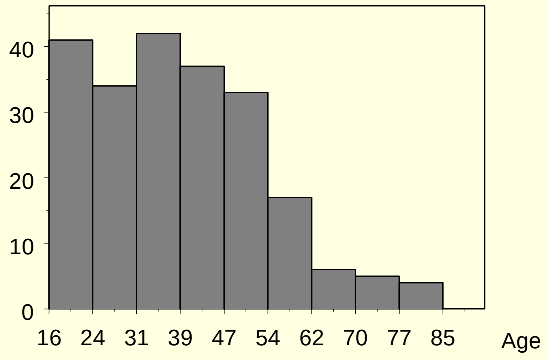
Bin	*Frequency*
20	2853
25	3709
30	4372
35	4366
40	4097
45	3588
50	2707
55	1831
60	1140
65	615
70	397
75	271
80	148
85	83
90	32
95	12
More	5

Histograms of Age Variable

Varying Window Size







Formula for Window Width

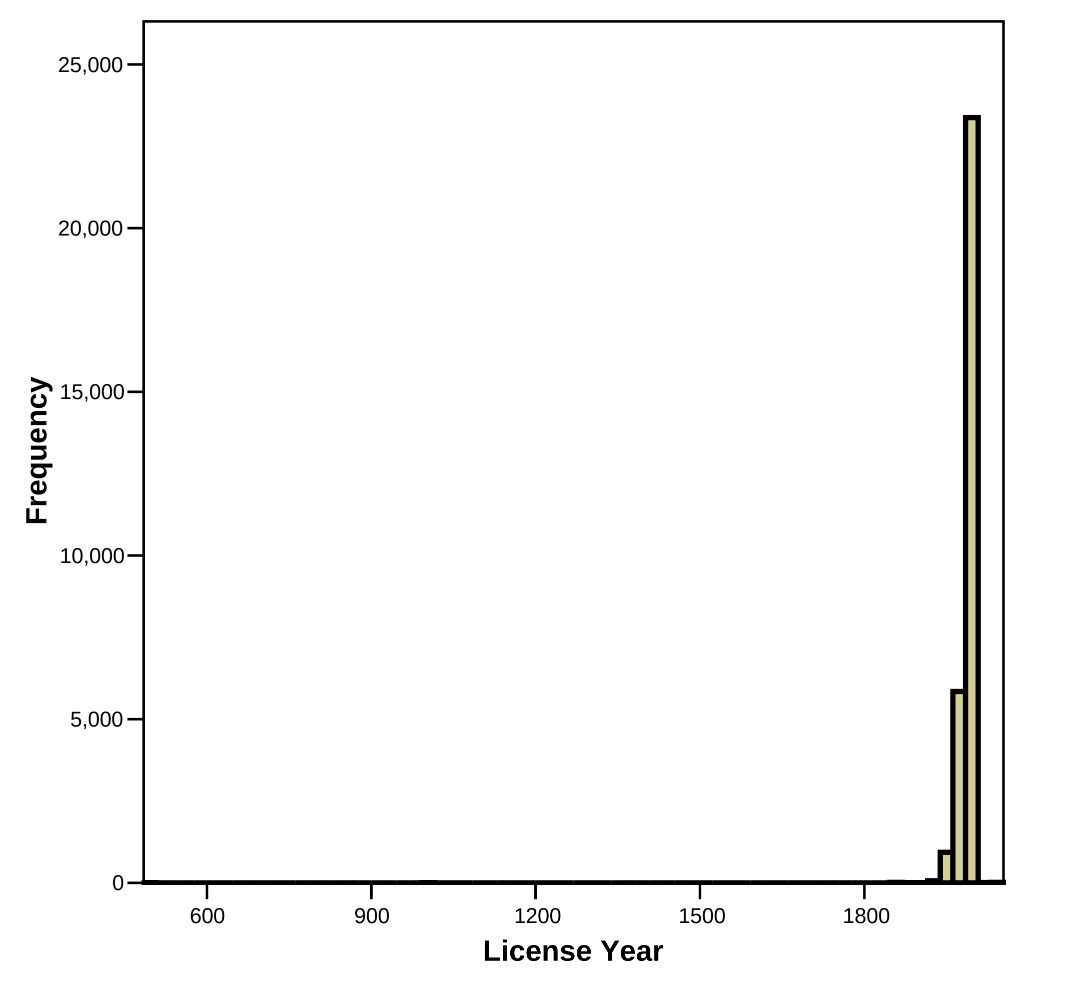
$$h = \frac{3.5\sigma}{\sqrt[\frac{1}{3}]{N}}$$

σ = standard deviation

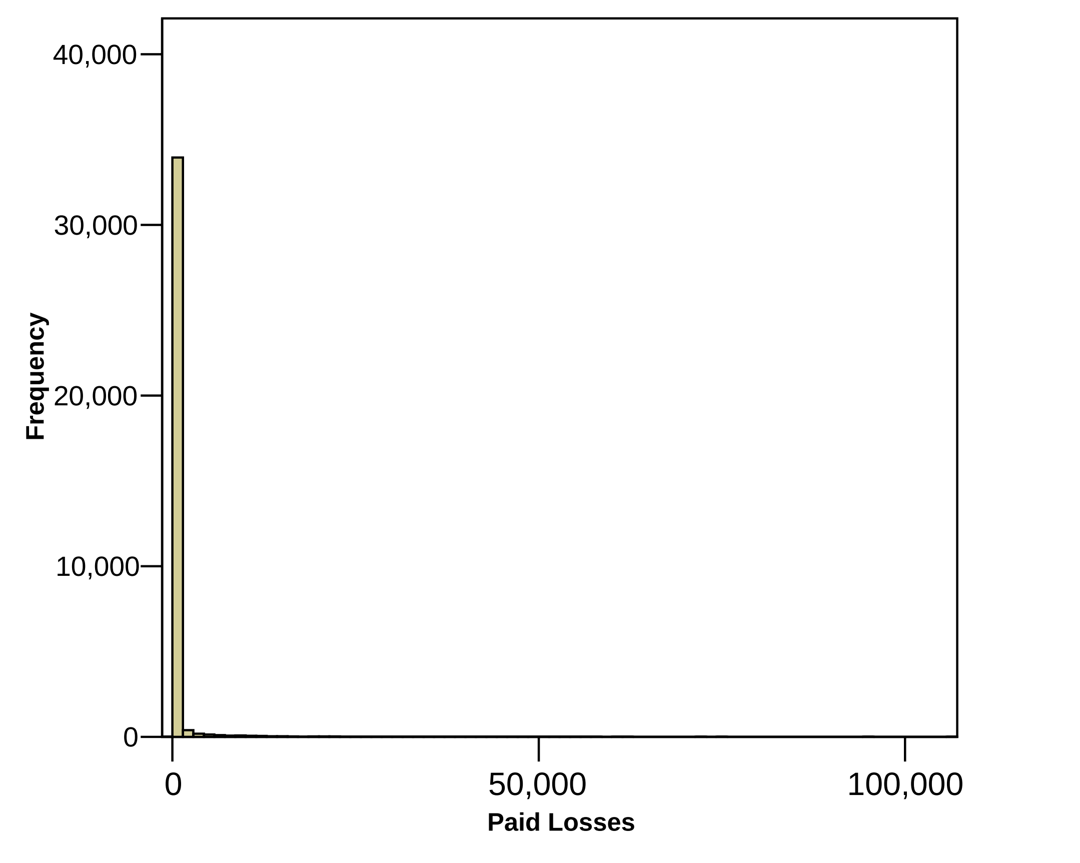
N=sample size

h =window width

Example of Suspicious Value

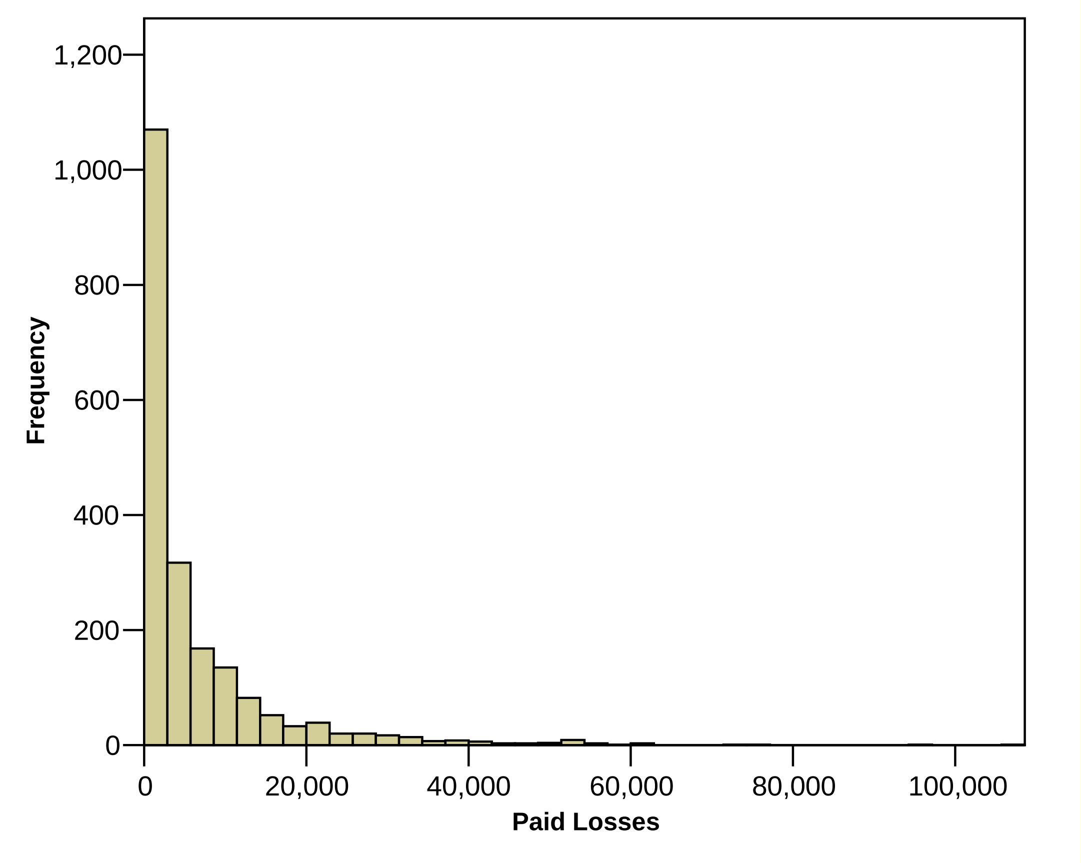


Discrete-Numeric Data



Filtered Data

Filter out Unwanted Records



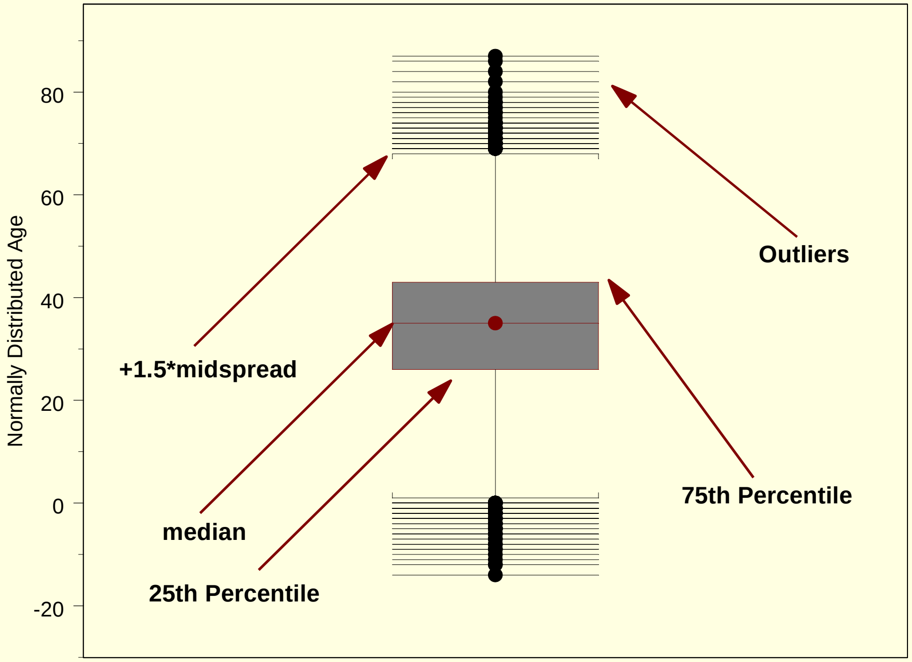
Box Plot Basics:
Five – Point Summary

- Minimum
- 1st quartile
- Median
- 2nd quartile
- Maximum

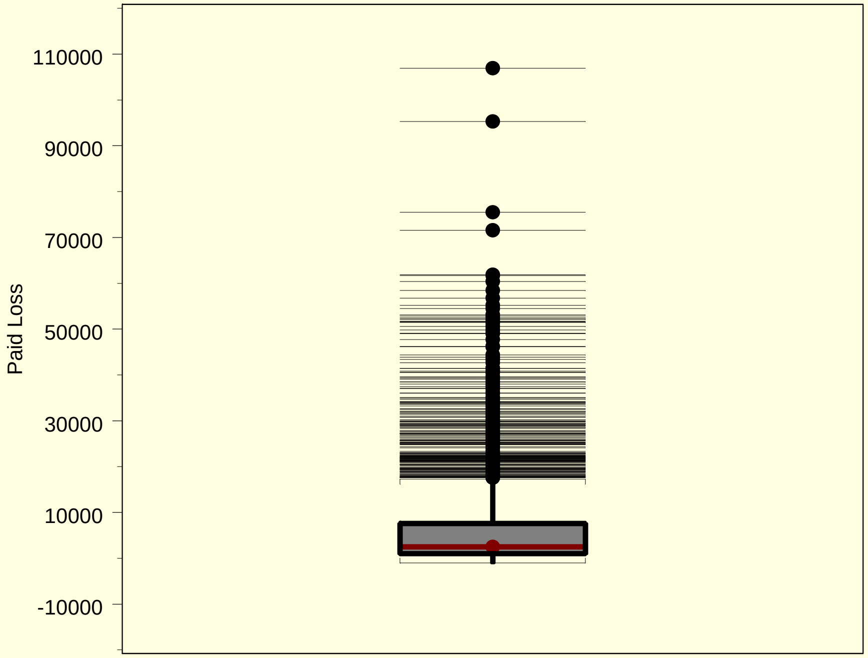
Functions for five point summary

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

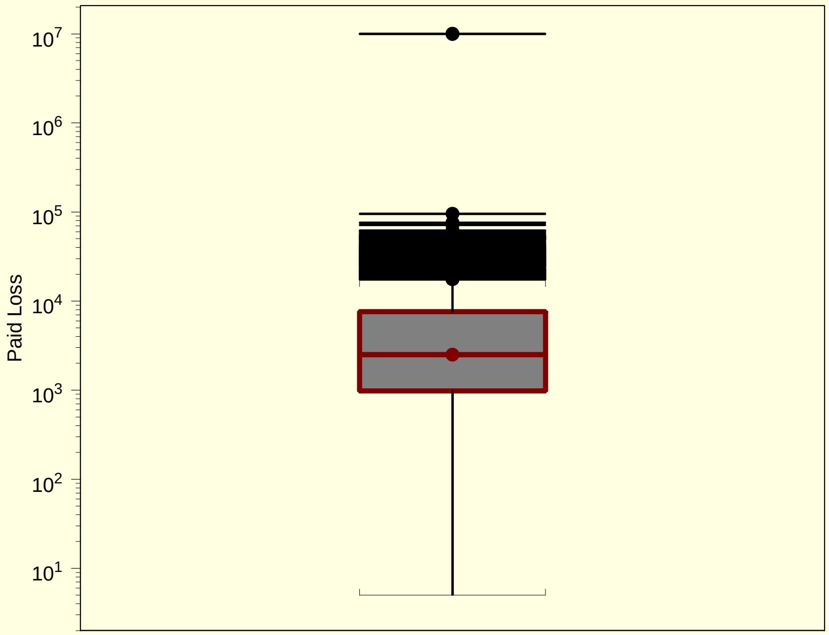
Box and Whisker Plot



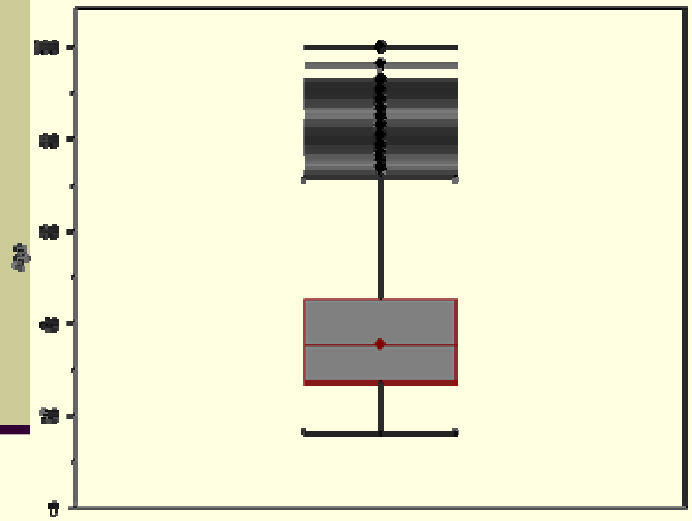
Plot of Heavy Tailed Data Paid Losses

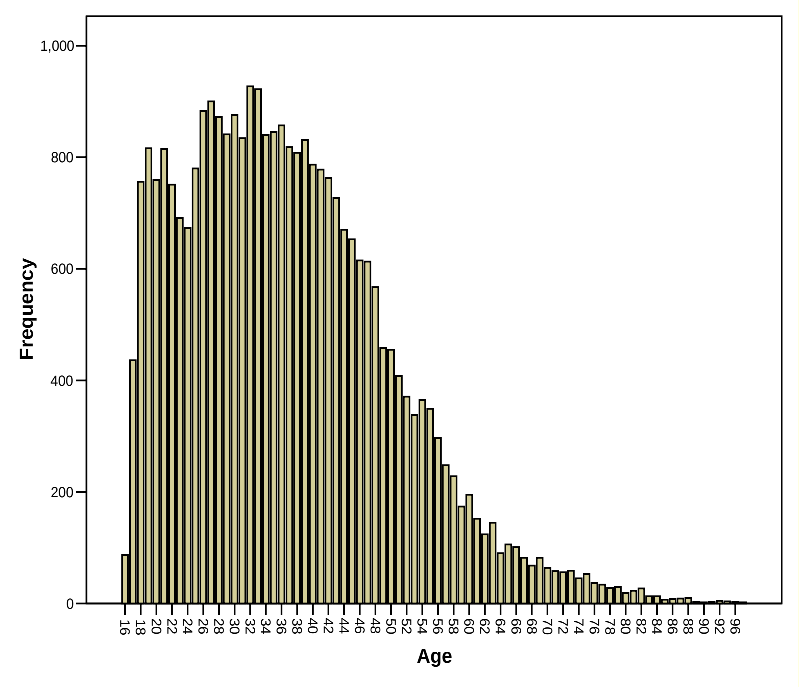


Heavy Tailed Data – Log Scale



Box and Whisker Example





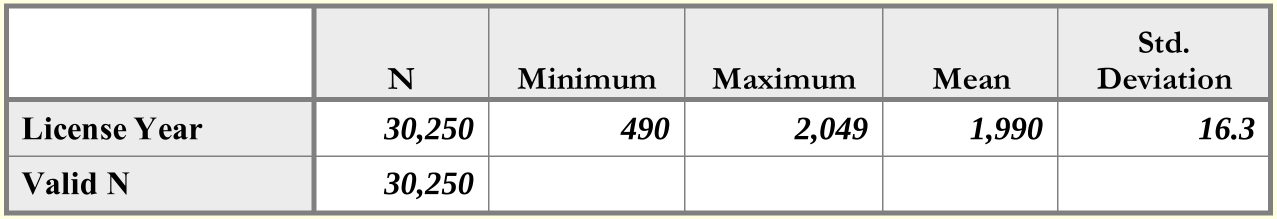
Descriptive Statistics

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***				

Data Spheres: The Mahalanobis Distance Statistic

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

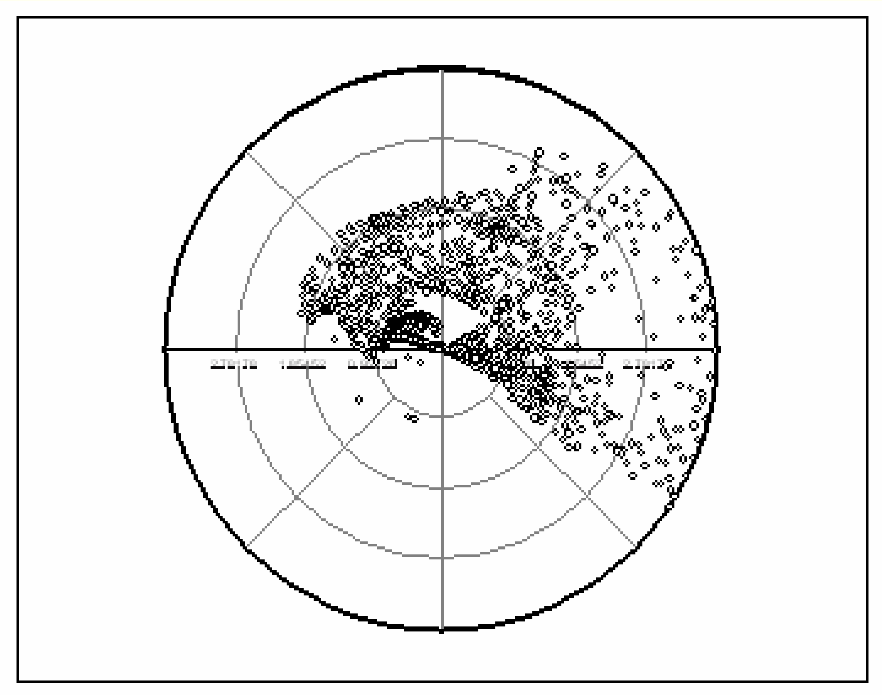
$\mathbf{x}$ is a vector of variables

$\boldsymbol{\mu}$ is a vector of means

$\boldsymbol{\Sigma}$ is a variance-covariance matrix

Screening Many Variables at Once

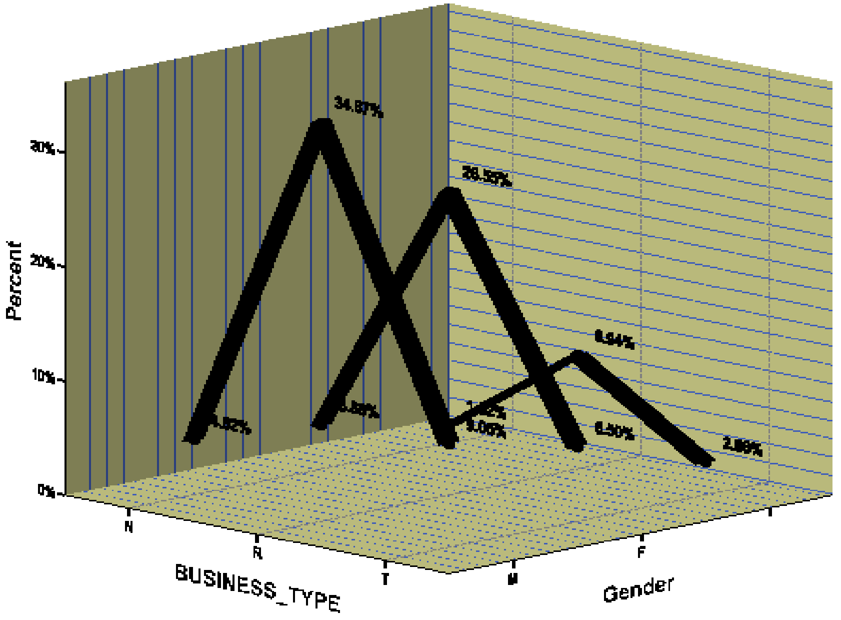
- Plot of Longitude and Latitude of zip codes in data
- Examination of outliers indicated drivers in Ca and PR even though policies only in one mid-Atlantic state



Records With Unusual Values Flagged

	Mahalanobis	Percentile of		License	Number	Number	Model	Incurred
Policy ID	Depth	Mahalanobis	Age	Year	of Cars	of Drivers	Year	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: Data Cubes



Categorical Data

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

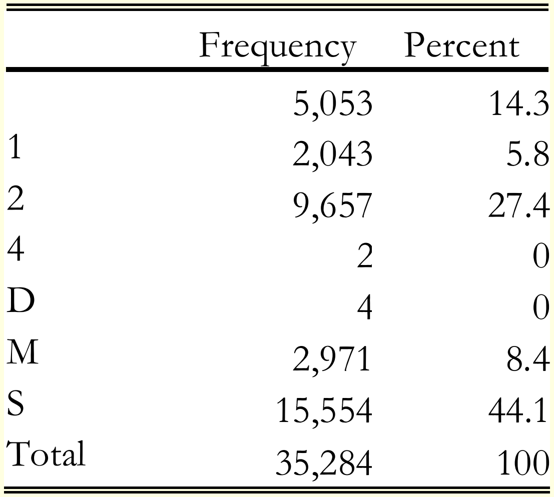
Gender

	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

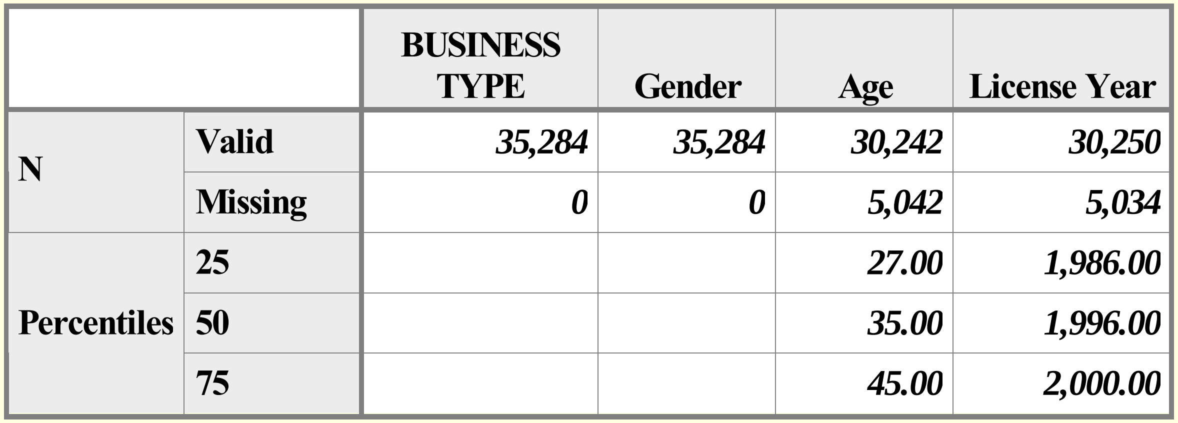
Marital Status



	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

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			

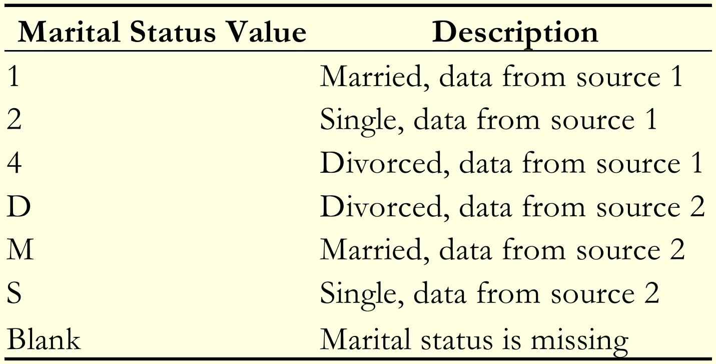
Missing Values

- 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

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 Claeaning Data”, CAS Winter Forum, March 2005, www.casact.org
- Find a comprehensive book for doing analysis in Excel such as: John Walkebach, *Excel 2003 Formulas* or Jospeh Schmuller, Statistical Analysis With Excel for Dummies
- If you use R, get a book like: Fox, John*, An R and S-PLUS Companion to Applied Regression*, Sage Publications, 2002