Raising Your Actuarial IQ
(Improving Information Quality)

CAS Data Management Educational Materials Working Party

Disclaimer

This presentation and the working party's other work products express the opinions of the members of the working party and not necessarily those of their employers or of the Casualty Actuarial Society.

Presenters

Moderator:

- **Virginia Prevosto**, FCAS

  Vice President, ISO

Panelists:

- **Aleksey Popelyukhin**, PhD

  Vice President, Information Systems

- **Keith Allen**, ACAS

  Vice President, Medical Mutual Liability Insurance Society of Maryland

AGENDA

- **Introduction**
- Data Life Cycle
- Data Management Best Practices
- Conclusions

2006 GIRO Data Quality Survey

- GIRO is the General Insurance Research Organisation; the property & casualty branch of the British actuarial profession
- Formed a working party to explore the impact of data quality on actuarial work and to make recommendations
- Working party's final report is "Dirty Data on Both Sides of the Pond" published in the Winter 2008 edition of the CAS *eForum*

2006 GIRO Data Quality Survey

- Working party conducted an informal survey in Britain, the U.S. and Canada
- Two questions:
  1. What percentage of time is spent on data quality issues?
  2. What proportion of projects are adversely affected by such issues?

Survey Conclusions

- Data quality issues have a significant impact on the work of general insurance (P&C) actuaries:
  - About a quarter of their time is spent on such issues
  - About a third of projects are adversely affected

“Actuarial IQ” Introduction

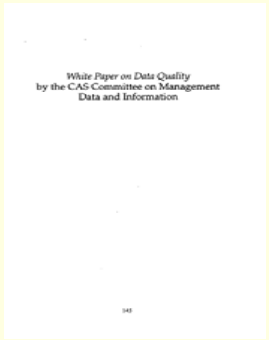
- Introduction to Data Quality and Data Management written by the CAS Data Management Educational Materials Working Party
- Directed at actuarial analysts as much as actuarial data managers:
  - what every actuary should know about data quality and data management
- “Information quality” because data quality is affected by processes as well as coding

AGENDA

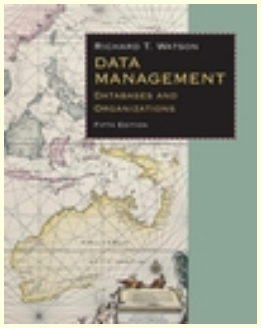
- Introduction
- **Data Life Cycle**
- Data Management Best Practices
- Conclusions

Principles of Data Quality: Perspectives

ASB – ASOP 23 – *"Data Quality"*



CAS Management Data and Information Committee: *"White Paper on Data Quality"*



Richard T. Watson
"Data Management: Databases and Organization"

Data Quality Fundamentals: ASOP No. 23

Due consideration to the following:

- **Appropriateness** for intended purpose
- **Reasonableness**
- **Comprehensiveness**
- Any known, material **limitations**
- The cost and feasibility of obtaining **alternative data**
- **The benefit** to be gained from an alternative data set
- **Sampling methods**

White Paper on Data Quality

Evaluating data quality consists of examining data for:

- **Validity**
- **Accuracy**
- **Reasonableness**
- **Completeness**

Watson

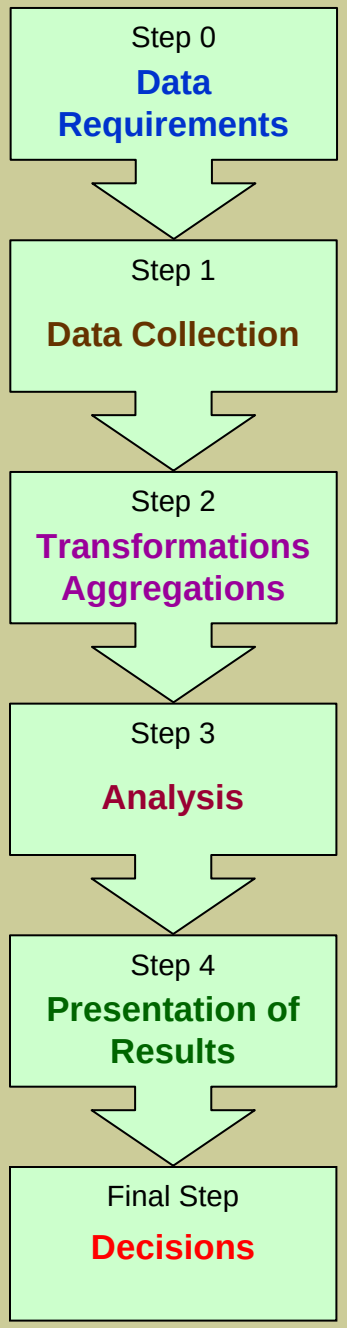
18 Dimensions of Data Quality:

- Many overlap with previously mentioned principles.
- Others describe ways of storing data
  e.g. Representational consistency, Precision
- Others go beyond data characteristics to processing and management

  e.g. Stewardship, Sharing, Timeliness, Interpretation

What is Data Quality?

- Quality data is data that is **appropriate** for its purpose.
- Quality is a **relative** not absolute concept.
  - Data for an annual rate study may not be appropriate for a class relativity analysis.
  - Promising predictor variables in Predictive Modeling may not have been coded or processed with that purpose in mind.



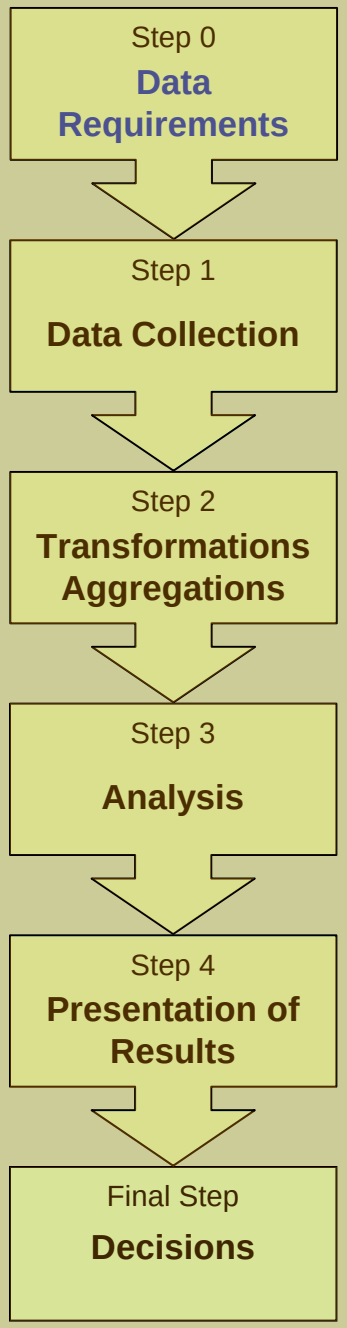
Data Flow

Information Quality involves **all** steps:

- Data Requirements
- Data Collection
- Transformations & Aggregations
- Actuarial Analysis
- Presentation of Results

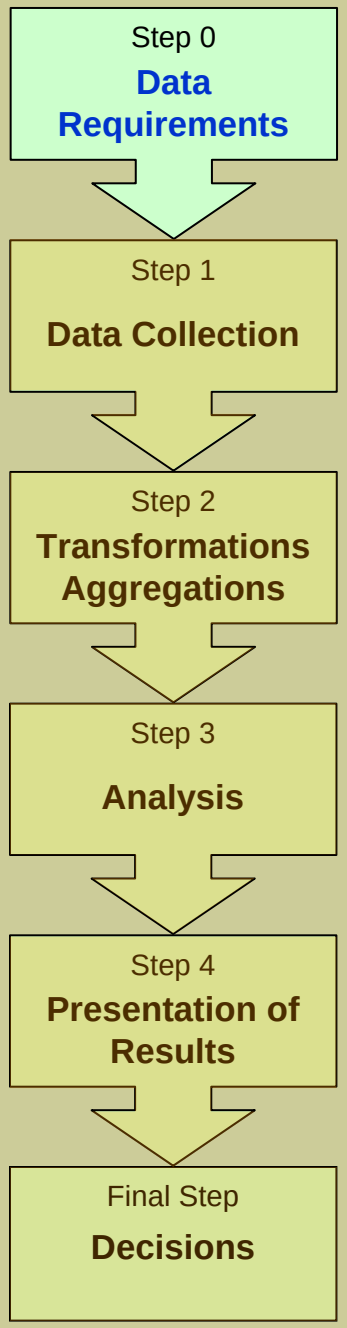
To improve Final Step:

- Making Decisions



Data Requirements

- Data managers know this step well
- Actuaries receive no formal training in Data Requirements concepts or theories
- This creates an opportunity for you to partner with your data managers
- One bridge that you can build with your data managers is to ask them to review the available **metadata** with you



Metadata

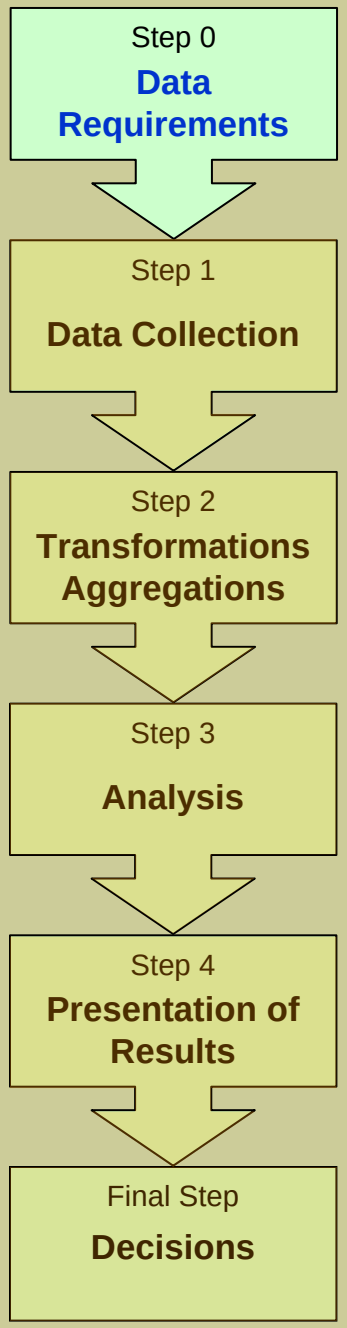
Big help in describing Data Requirements – **Metadata!**

- Data that Describes the Data
  - Key Data Management Tool
  - Reduces Risky Assumptions

    E.g., does CWP mean…

    Closed with Payment?

    Closed without Payment?



Example – Marital Status

- What is in the Marital Status Variable?

Marital Status

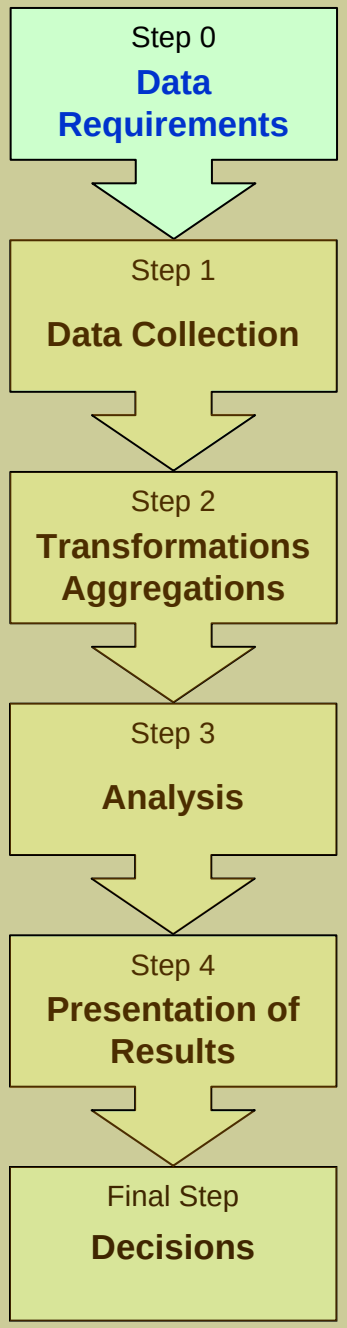
	Frequency	Percent
	5,053	14.3
1	2,043	5.8
2	9,657	27.4
4	2	0
D	4	0
M	2,971	8.4
S	15,554	44.1
Total	35,284	100

Single?

Married?

Polygamist?

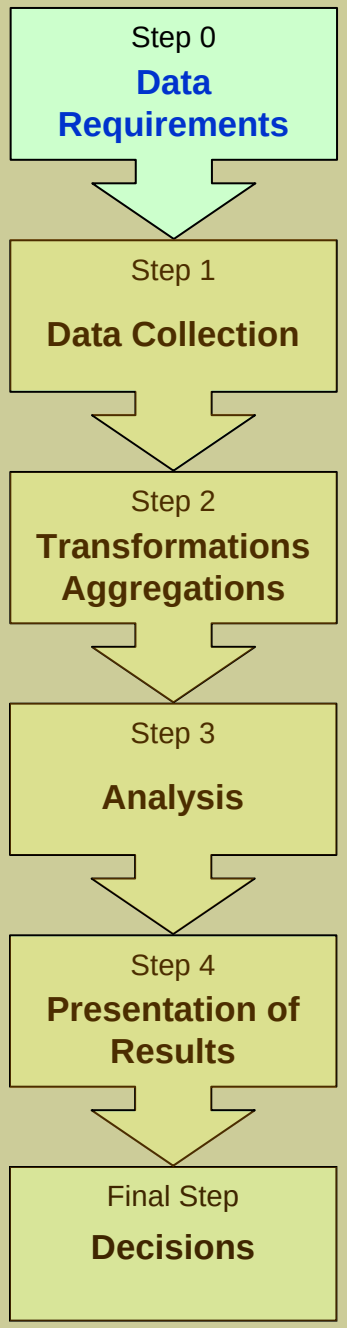
Single / Separated?



Example: What is the Marital Status Variable?

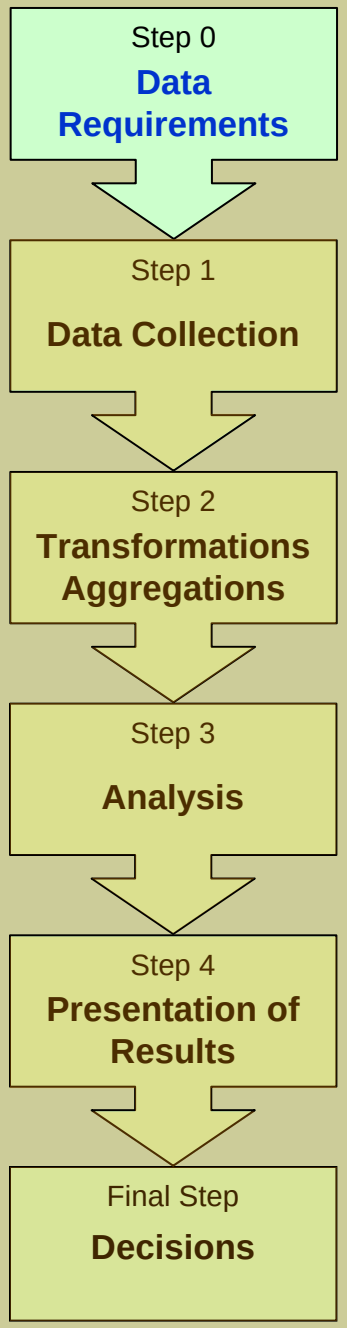
Example of Metadata

Marital Status Value	Description
1	Married, data from source 1, straight move of field ms_code
2	Single, data from source 1, straight move of field ms_code
4	Divorced, data from source 1, straight move of field ms_code
D	Divorced, data from source 2, straight move of mstatus
M	Married, data from source 2, straight move of mstatus
S	Single, data from source 2, straight move of mstatus
Blank	Marital status is missing



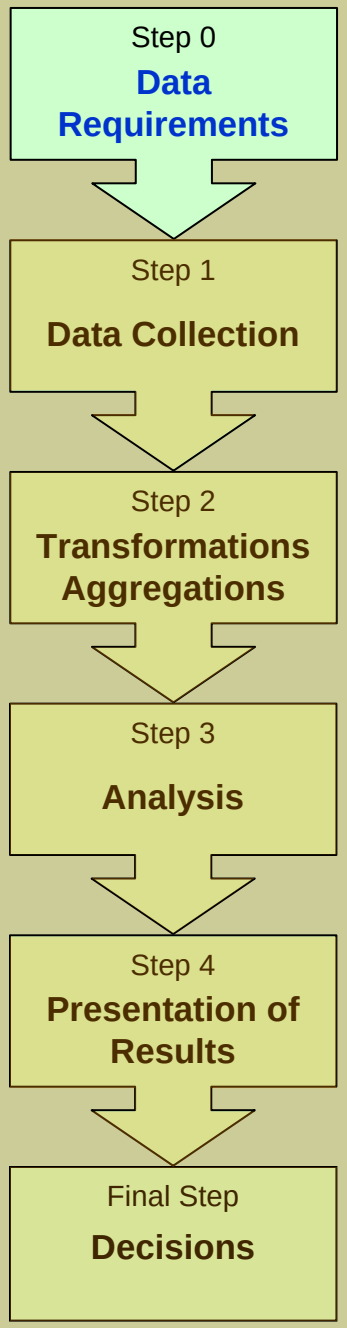
What is in Metadata?

- Business Rules
- Data Processing Rules
- Report Compilation and Extraction Process
- Other



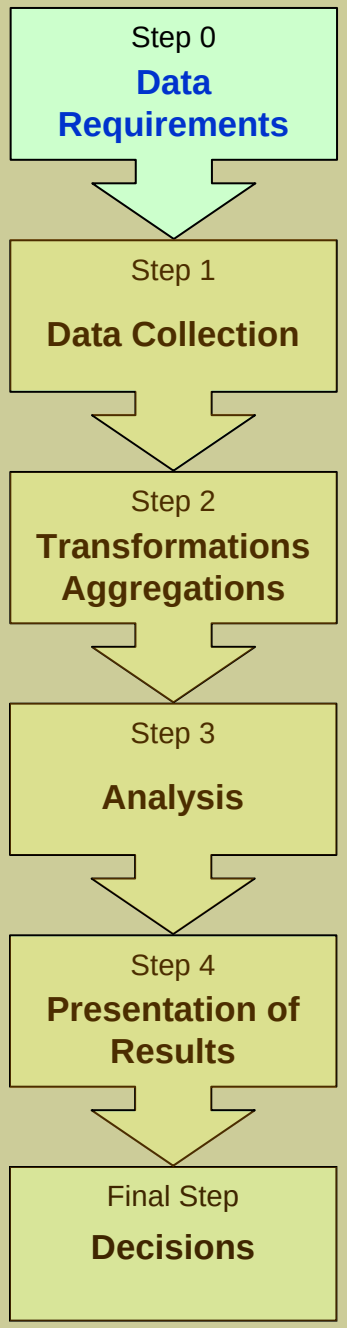
What is in Metadata?

- Business Rules
  - Data Elements
    - Definition of Field, e.g.,
      - How Claims are Defined
      - How Exposure is Calculated
    - Format of Field
      - mm/dd/yyyy
      - #,##0.00
    - Valid Values and Interdependencies
      - Alpha Only
      - Driver = Yes and Age > 15



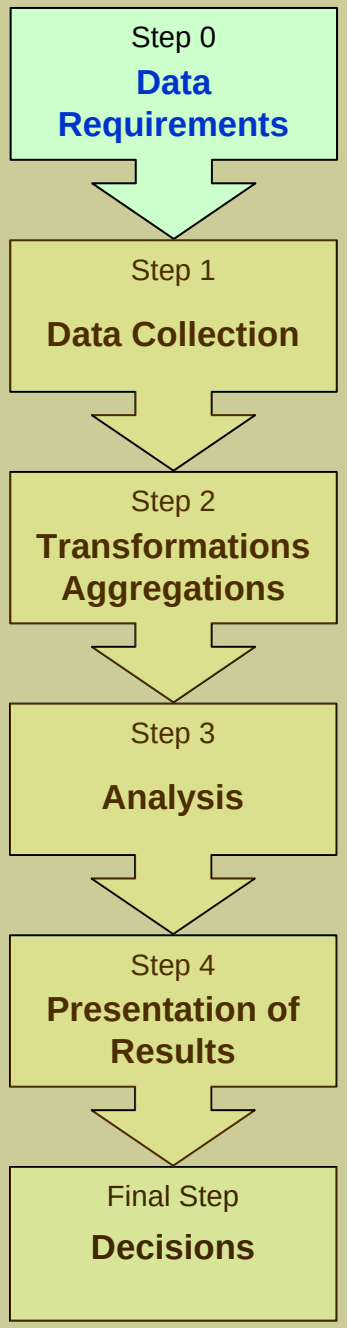
What is in Metadata?

- Data Processing Rules
  - How Database is Populated
  - Sources of Data
  - Handling of Missing Data



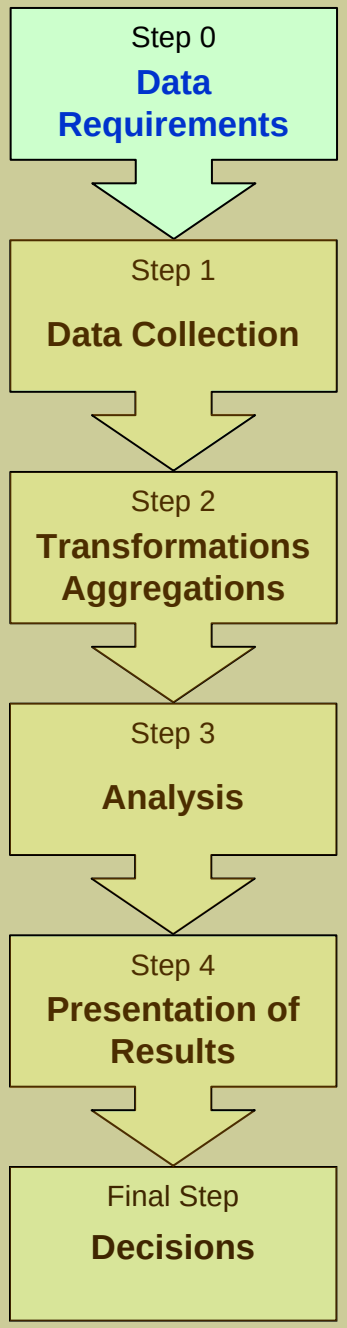
What is in Metadata?

- Report Compilation and Extraction Process
  - How Data is Selected or Bypassed
  - Fiscal Period
  - Accounting Date for Transactions
  - Actuarial Evaluation Date
  - Calculations
  - Mappings



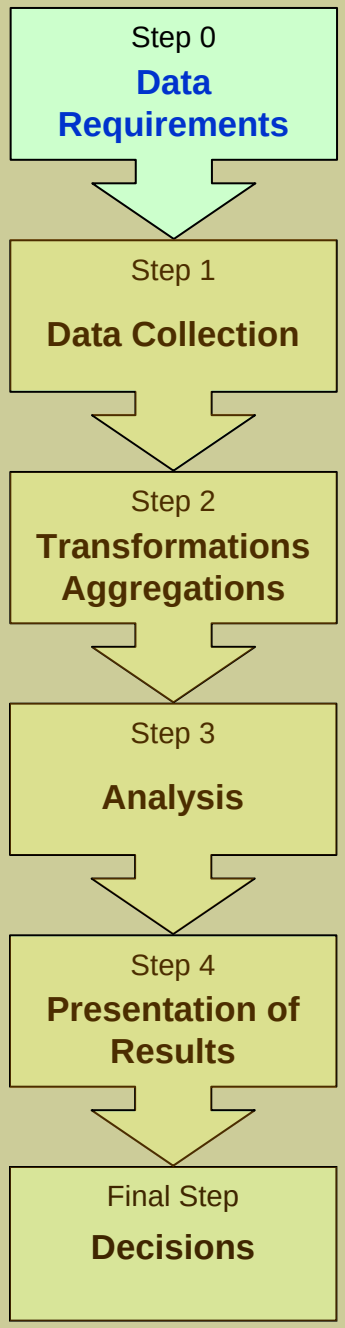
What is in Metadata?

- Other
  - Process Flow Documentation
  - Versioning



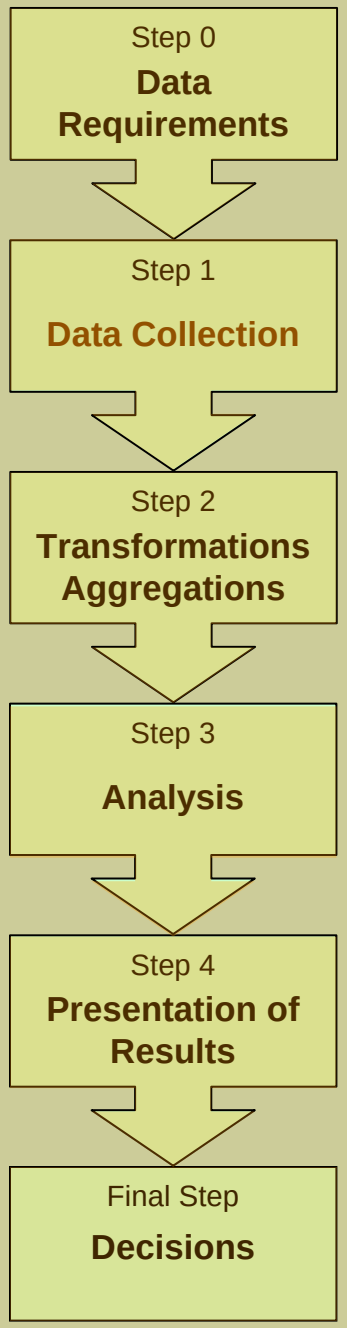
Why Actuaries Need Metadata

- Can result in better analyses
- Can avoid being misinformed about data or what it represents
- Can identify if anything changed during the experience period
  - But only if
    - They ask to receive this
    - Actually compare metadata lists / files



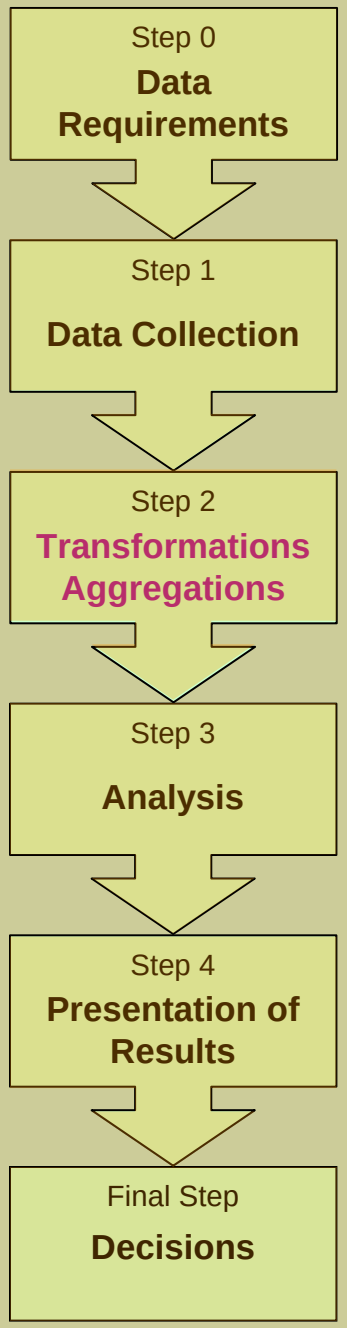
Example of Metadata

- Statistical Plans in the Property Casualty Insurance Industry
  - General Reporting Requirements
  - Data Element Definitions
  - Standardize Data to the Extent Possible



Data Collection

- Data supplier management
  - Let suppliers know what you want
  - Provide feedback to suppliers
  - Balance the following
    - Known issues with supplier
    - Importance to the business
    - Supplier willingness to experiment together
    - Ease of meeting face to face



Transformations and Aggregations

- In this step data are put into standardized structures and then combined into larger, more centralized data sets
- “Actuarial IQ” introduces two ways to improve IQ in this step:
  - Exploratory Data Analysis (EDA)
  - Data Audits

Step 0
Data Requirements

Step 1
Data Collection

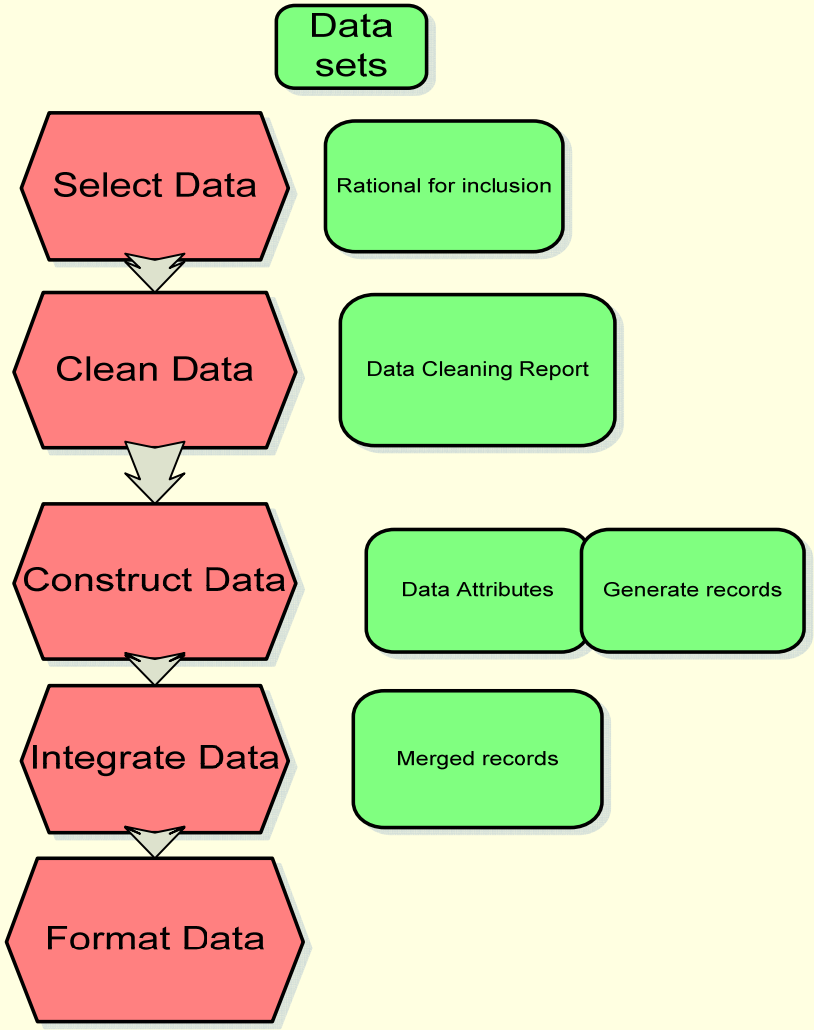
Step 2
Transformations Aggregations

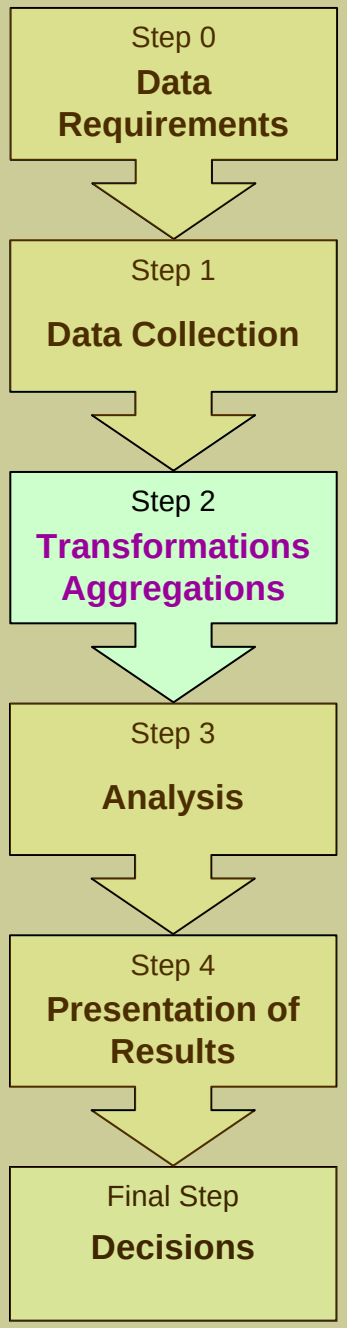
Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Data Preprocessing





EDA: Overview

- Typically first step in analyzing data
- Purpose:
  - Explore structure of the data
  - Find outliers and errors
- Uses simple statistics and graphical techniques
- Examples include histograms, descriptive statistics and frequency tables

Step 0
Data Requirements

Step 1
Data Collection

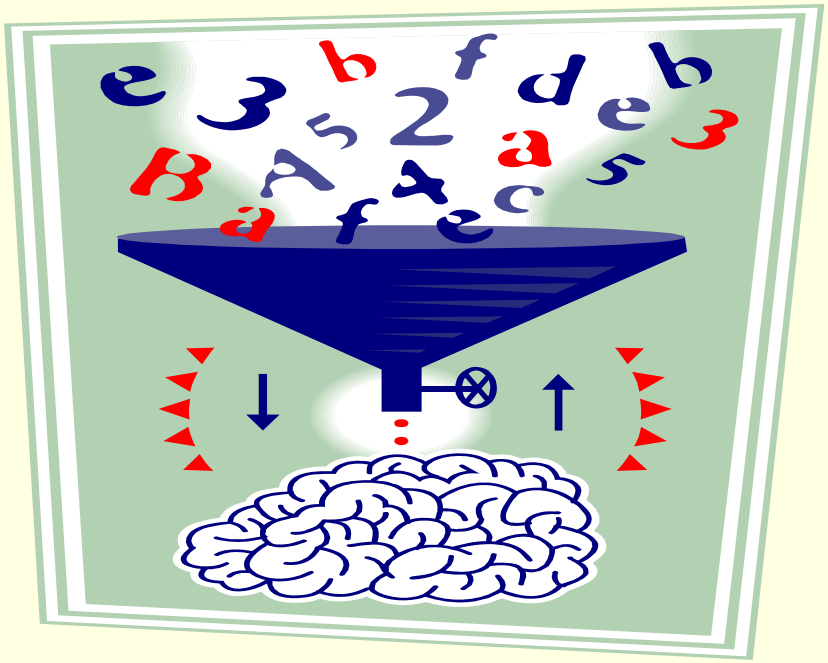
Step 2
Transformations
Aggregations

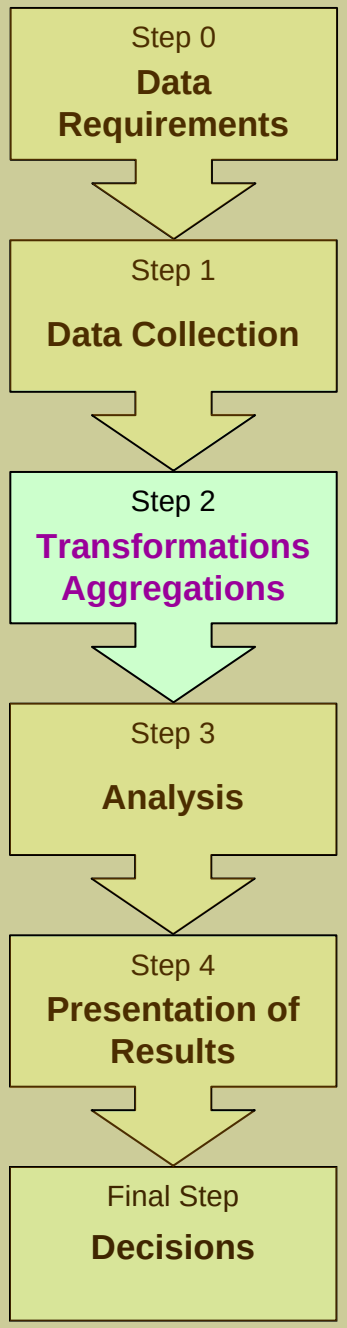
Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

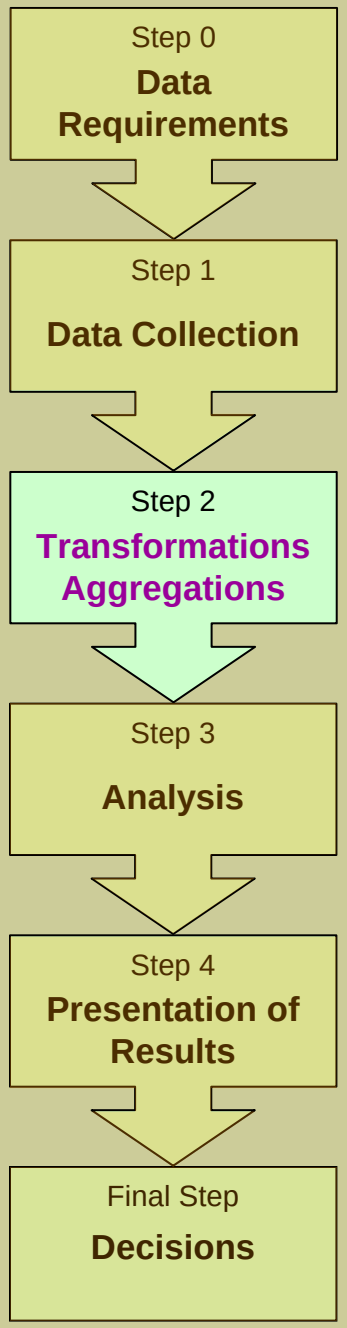
EDA: Working Example





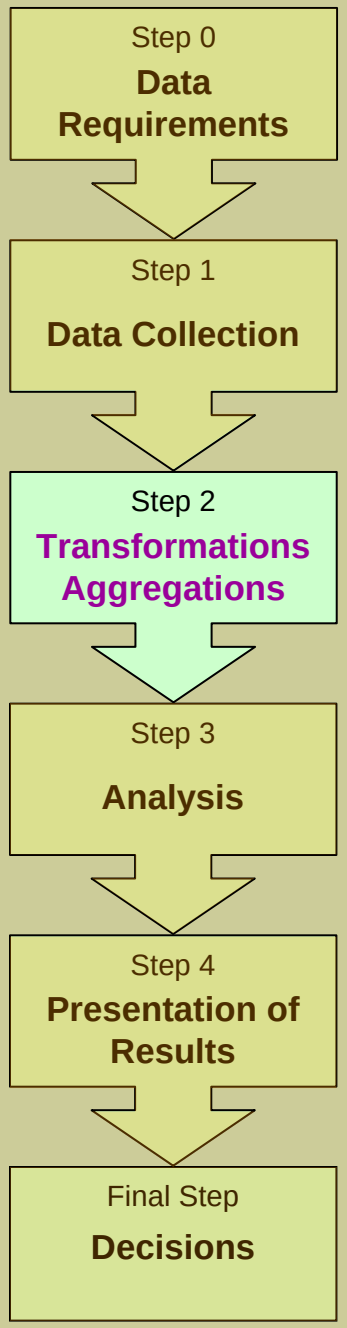
EDA: Working Example

- Tests ➔Data Quality
  - Validity
  - Accuracy
  - Reasonableness
  - Completeness
- In-force Premium File for Personal Auto
  - One record per insured vehicle
- Can use Microsoft Excel for Small Data Sets



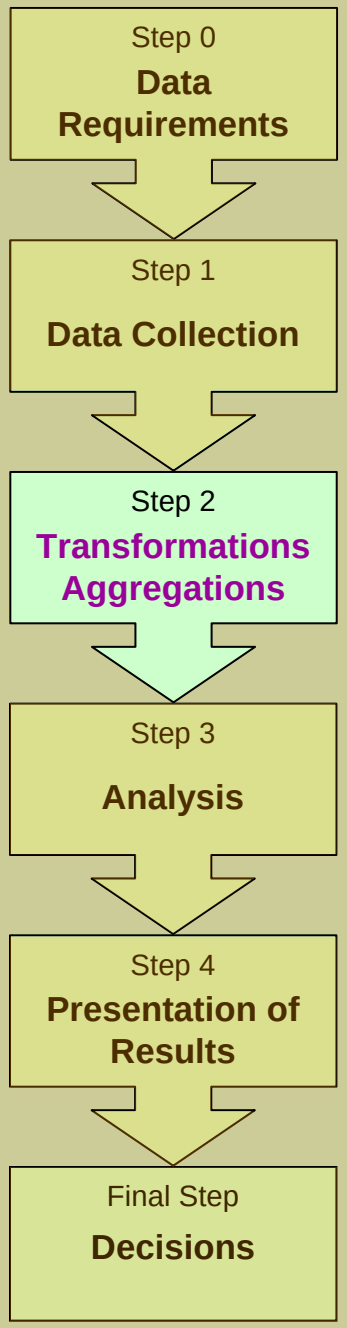
EDA: Working Example

1. Data Receipt
   - File already had field headings and was parsed
   - Check for "personally identifiable" information and mask if necessary



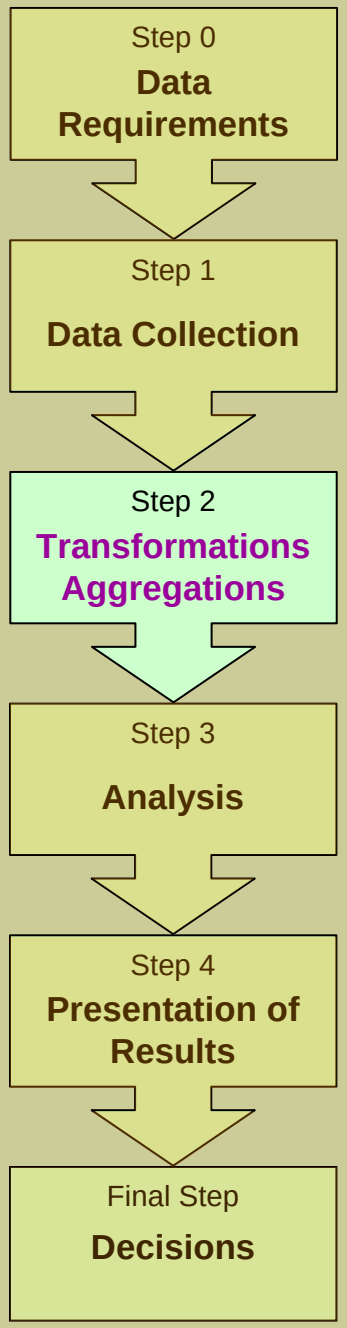
EDA: Working Example

2. Initial Preparation
   - Add Record ID for each row in data set
   - Insert column number
   - Compile list of data elements
   - Add derived fields such as ZIP Code, County, Term
   - Reformat fields (if necessary)
   - Generate control totals (counts and dollars)



EDA: Working Example

3. Explore Structure of the Data
   - Validity
   - Accuracy
   - Reasonableness
   - Completeness
   - Maintain list of any corrections made to data



EDA: Working Example

4. Final Preparation
   - Delete fields not relevant to analysis, e.g., name of insured, address fields

5. Analysis
   - May need to cycle back as results emerge

Step 0
Data Requirements

Step 1
Data Collection

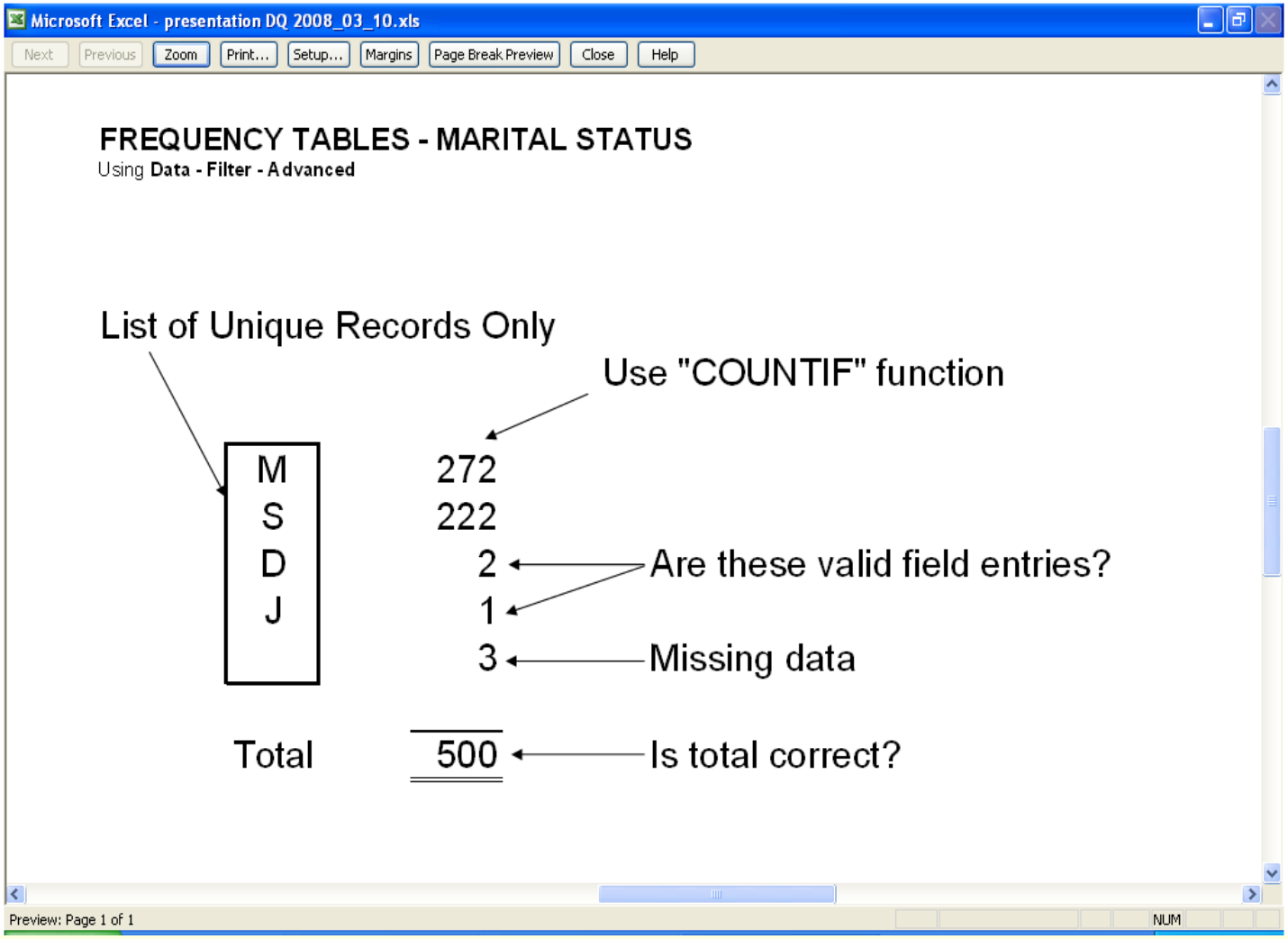
Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Frequency Table



Step 0
Data Requirements

Step 1
Data Collection

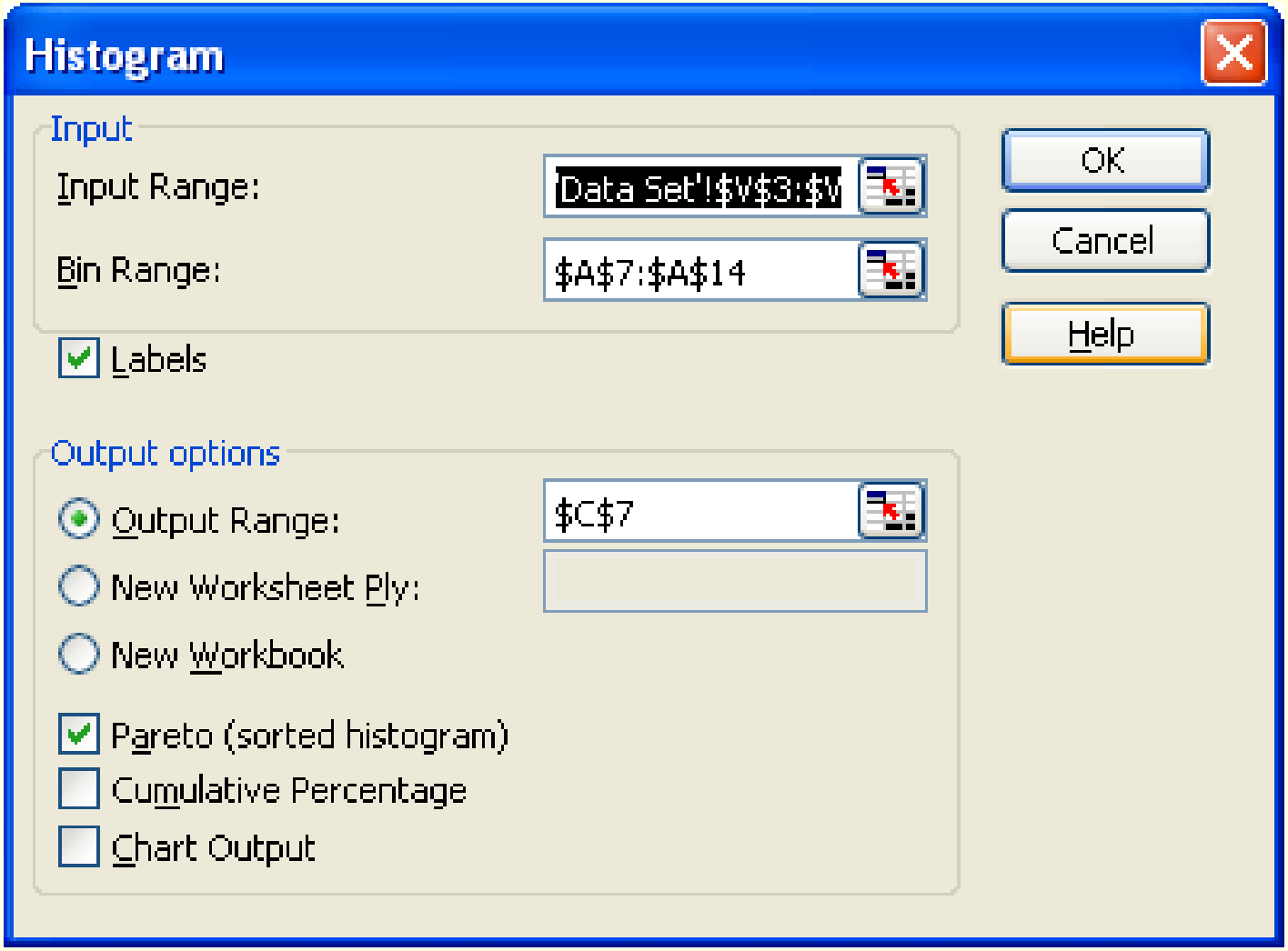
Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Histograms



Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Histograms

Microsoft Excel - presentation DQ 2008_03_10.xls

Next | Previous | Zoom | Print... | Setup... | Margins | Page Break Preview | Close | Help

HISTOGRAM ON BI PREMIUM

Under **Tools - Data Analysis**

Premium

100
200
300
400
700
1000
1300

Bin Range

PARETO (sorted histogram)

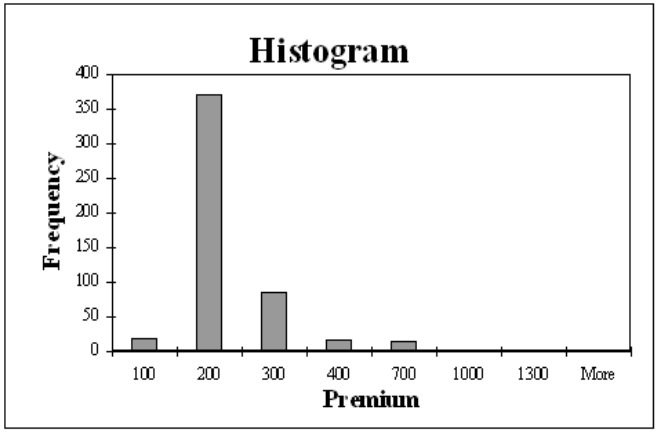
Premium	*Frequency*	*Premium*	*Frequency*
100	17	200	370
200	370	300	84
300	84	100	17
400	15	400	15
700	13	700	13
1000	0	1300	1
1300	1	1000	0
More	0	More	0

CUMULATIVE PERCENTAGE

Premium	*Frequency*	*Cumulative %*
100	17	3.40%
200	370	77.40%
300	84	94.20%
400	15	97.20%
700	13	99.80%
1000	0	99.80%
1300	1	100.00%
More	0	100.00%

CHART OUTPUT

Premium	*Frequency*
100	17
200	370
300	84
400	15
700	13
1000	0
1300	1
More	0



Preview: Page 1 of 1

NUM

Step 0
Data Requirements

Step 1
Data Collection

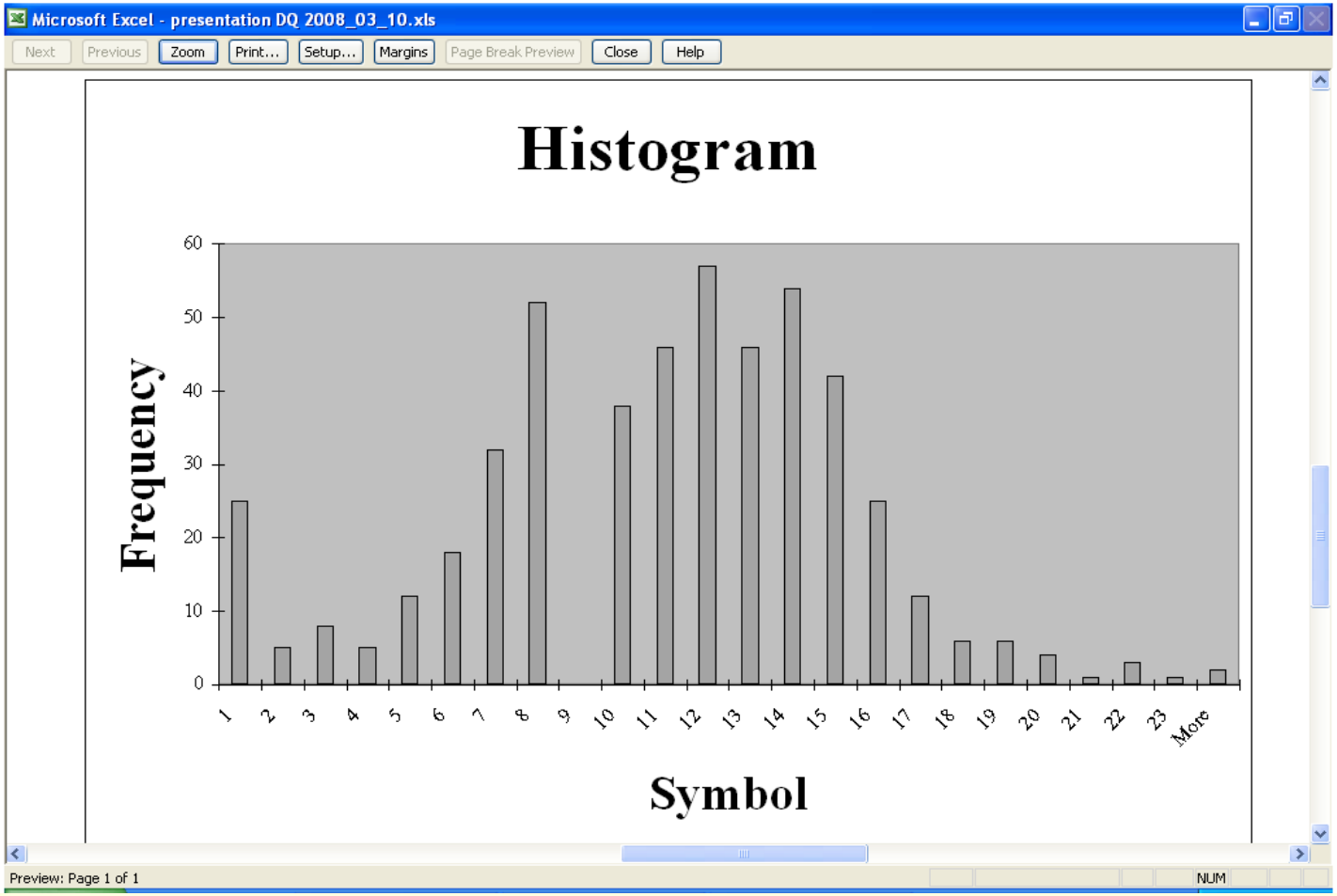
Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Histograms



Step 0
Data Requirements

Step 1
Data Collection

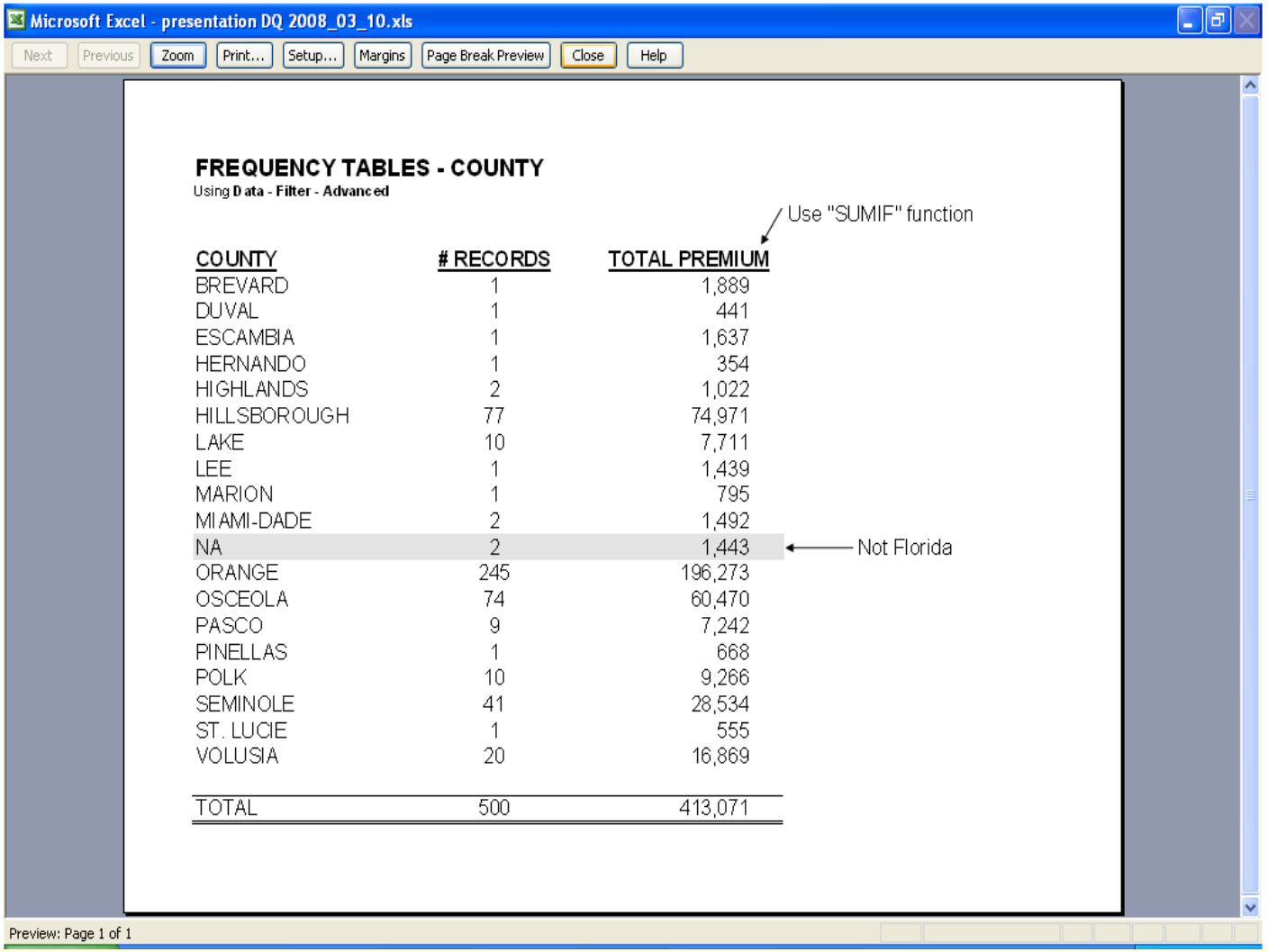
Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Frequency Tables



FREQUENCY TABLES - COUNTY

Using Data - Filter - Advanced

Use "SUMIF" function

COUNTY	# RECORDS	TOTAL PREMIUM
BREVARD	1	1,889
DUVAL	1	441
ESCAMBIA	1	1,637
HERNANDO	1	354
HIGHLANDS	2	1,022
HILLSBOROUGH	77	74,971
LAKE	10	7,711
LEE	1	1,439
MARION	1	795
MIAMI-DADE	2	1,492
NA	2	1,443
ORANGE	245	196,273
OSCEOLA	74	60,470
PASCO	9	7,242
PINELLAS	1	668
POLK	10	9,266
SEMINOLE	41	28,534
ST. LUCIE	1	555
VOLUSIA	20	16,869
TOTAL	500	413,071

← Not Florida

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Frequency Tables

Microsoft Excel - presentation DQ 2008_03_10.xls

Next | Previous | Zoom | Print... | Setup... | Margins | Page Break Preview | Close | Help

FREQUENCY TABLES - BI LIMIT

Using **Data - Pivot Tables**

BI Limit	Data	Total	Avg Prem	Freq %
$10/$20	Count of RecID	437		
	Sum of BI Premium	71,554	164	0.874
$15/$30	Count of RecID	4		
	Sum of BI Premium	963	241	0.008
$20/$40	Count of RecID	3		
	Sum of BI Premium	968	323	0.006
$25/$50	Count of RecID	21		
	Sum of BI Premium	4,590	219	0.042
$50/$100	Count of RecID	4		
	Sum of BI Premium	1,072	268	0.008
$100/$300	Count of RecID	31		
	Sum of BI Premium	9,705	313	0.062
Total Count of RecID		500		
Total Sum of BI Premium		88,852	178	1.000

Preview: Page 1 of 1 NUM

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Frequency Tables

Microsoft Excel - presentation DQ 2008_03_10.xls

Next | Previous | Zoom | Print... | Setup... | Margins | Page Break Preview | Close | Help

FREQUENCY TABLES - THREE WAY

Using **Data - Pivot Tables**

Count of RecID		Oper Good Student		
Oper Mature Operator	Decade	NO	YES	Grand Total
NO	1	10		10
	2	74	2	76
	3	125		125
	4	128		128
	5	94		94
	6	37		37
	7	13	1	14
	8	2		2
NO Total		483	3	486
YES	2	1		1
	5	2		2
	6	5		5
	7	5		5
	8	1		1
YES Total		14		14
Grand Total		497	3	500

Preview: Page 1 of 1 NUM

Step 0
Data Requirements

Step 1
Data Collection

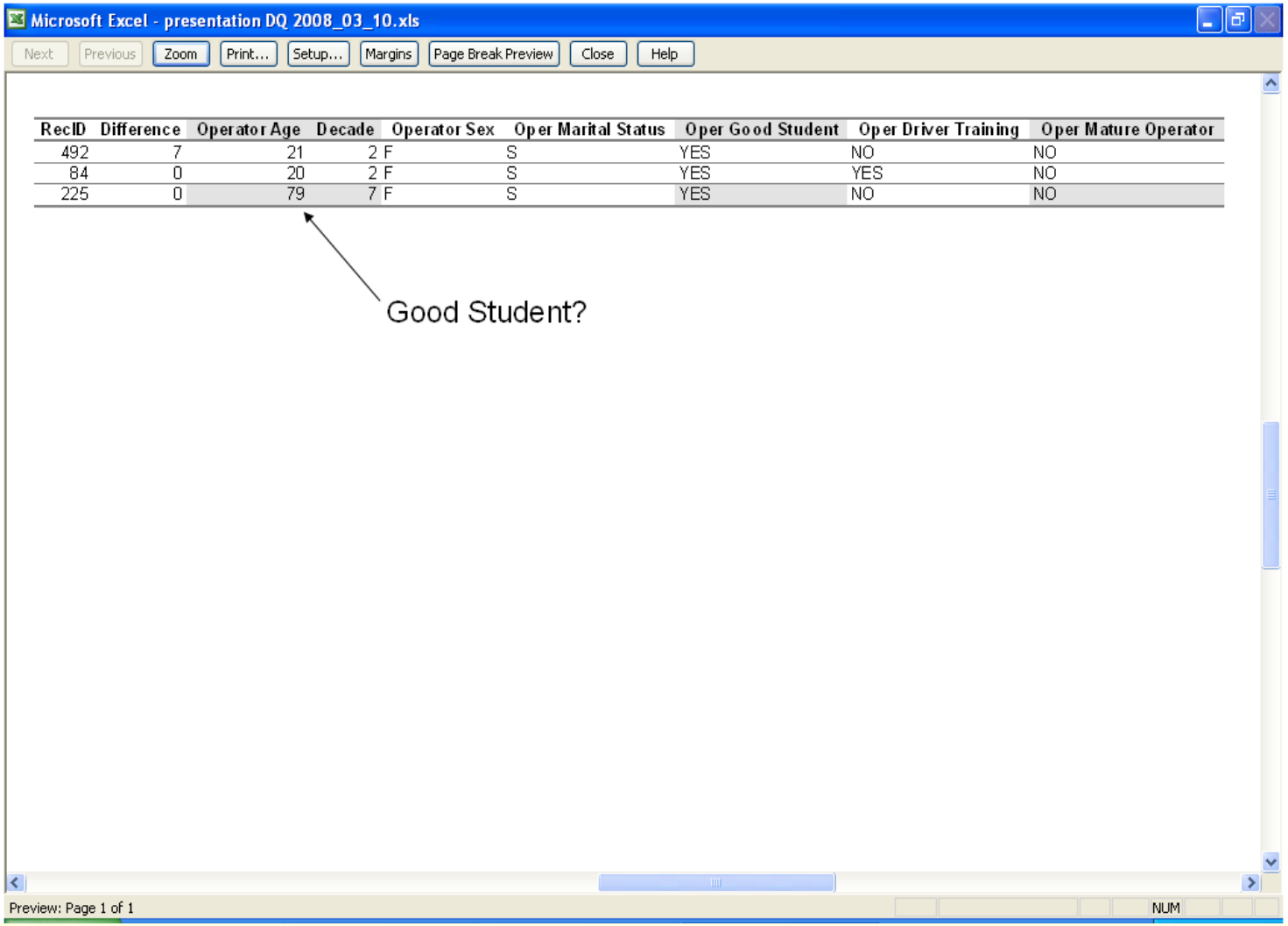
Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Frequency Tables



RecID	Difference	Operator Age	Decade	Operator Sex	Oper Marital Status	Oper Good Student	Oper Driver Training	Oper Mature Operator
492	7	21	2	F	S	YES	NO	NO
84	0	20	2	F	S	YES	YES	NO
225	0	79	7	F	S	YES	NO	NO

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Descriptive Statistics

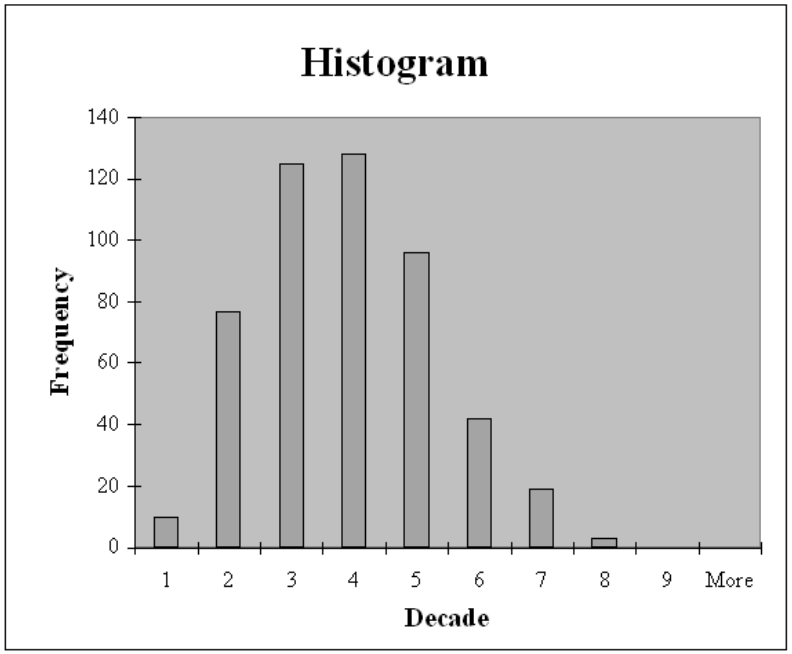
Microsoft Excel - presentation DQ 2008_03_10.xls

Next | Previous | Zoom | Print... | Setup... | Margins | Page Break Preview | Close | Help

DESCRIPTIVE STATISTICS - OPERATOR AGE

Under **Tools - Data Analysis**

Operator Age	
Mean	43.4
Standard Error	0.6
Median	42.0
Mode	48.0
Standard Deviation	13.9
Sample Variance	192.1
Kurtosis	(0.4)
Skewness	0.4
Range	67
Minimum	17
Maximum	84
Sum	21,702
Count	500
Largest(5)	79
Smallest(5)	18



Preview: Page 1 of 1

NUM

Step 0
Data Requirements

Step 1
Data Collection

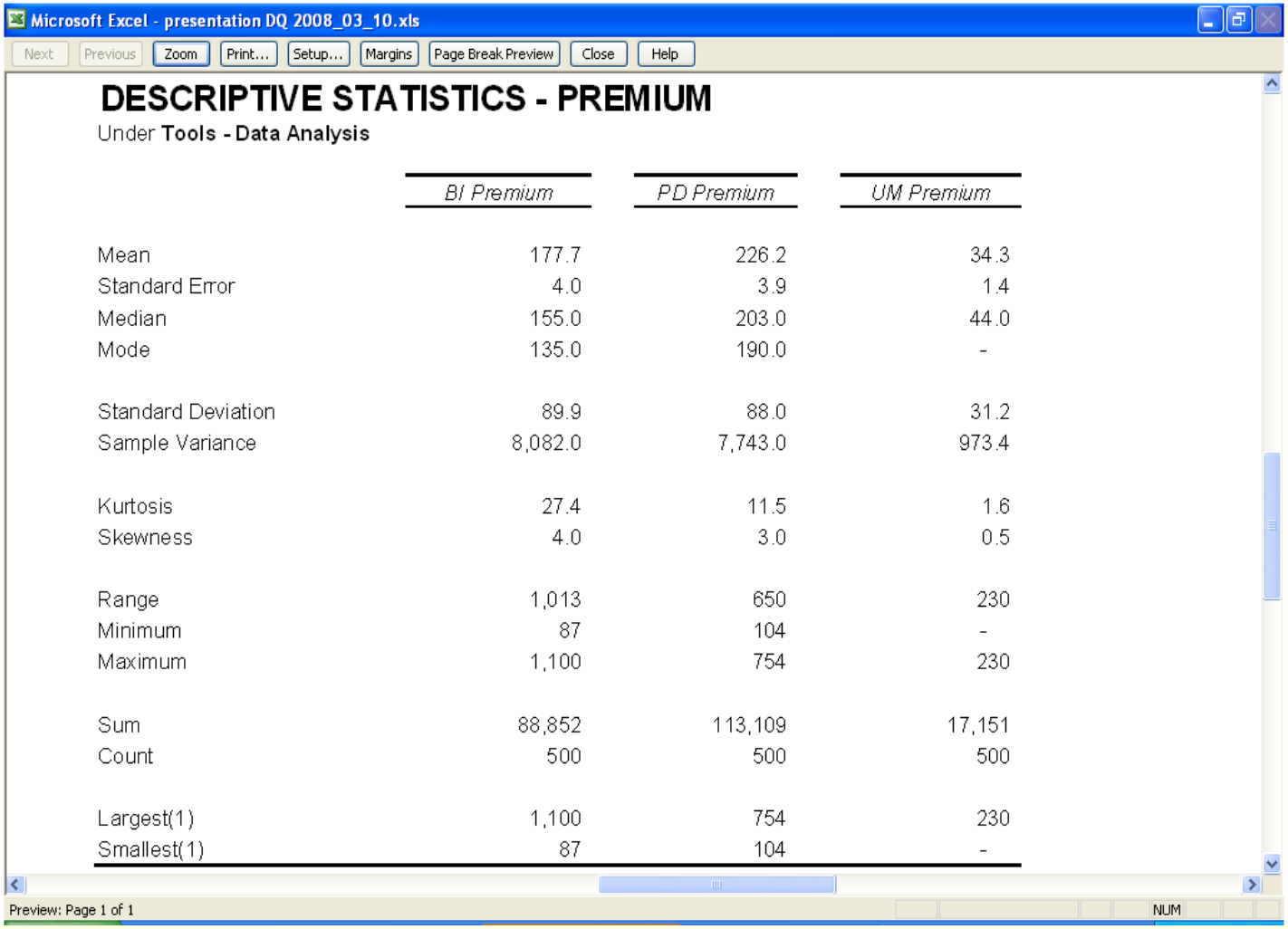
Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

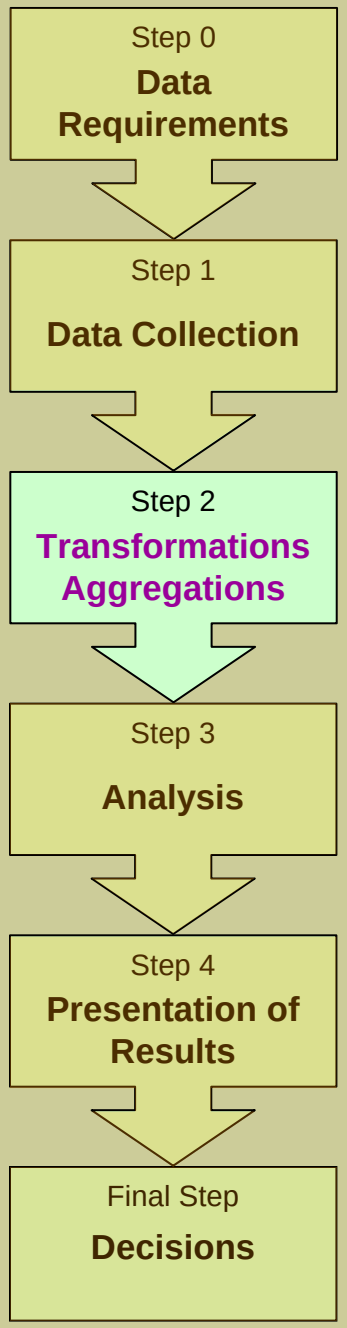
EDA: Descriptive Statistics



DESCRIPTIVE STATISTICS - PREMIUM

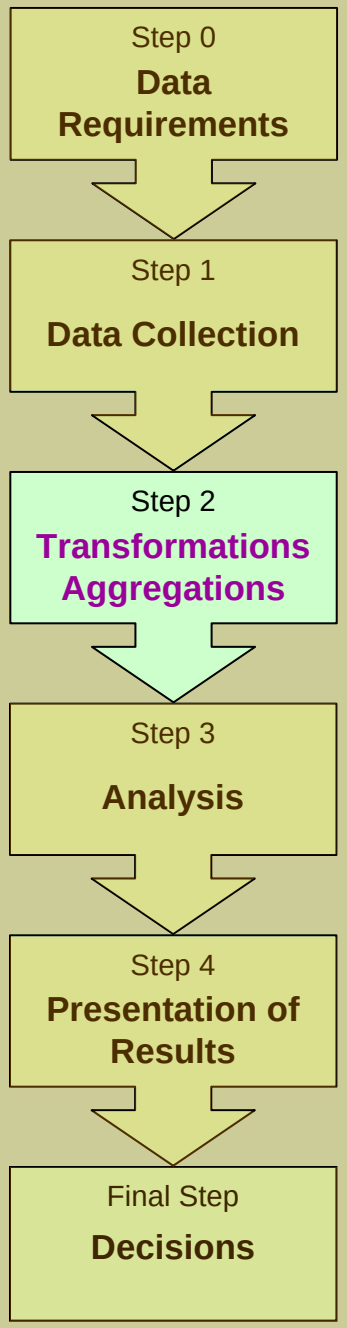
Under **Tools - Data Analysis**

	BI Premium	*PD Premium*	*UM Premium*
Mean	177.7	226.2	34.3
Standard Error	4.0	3.9	1.4
Median	155.0	203.0	44.0
Mode	135.0	190.0	-
Standard Deviation	89.9	88.0	31.2
Sample Variance	8,082.0	7,743.0	973.4
Kurtosis	27.4	11.5	1.6
Skewness	4.0	3.0	0.5
Range	1,013	650	230
Minimum	87	104	-
Maximum	1,100	754	230
Sum	88,852	113,109	17,151
Count	500	500	500
Largest(1)	1,100	754	230
Smallest(1)	87	104	-



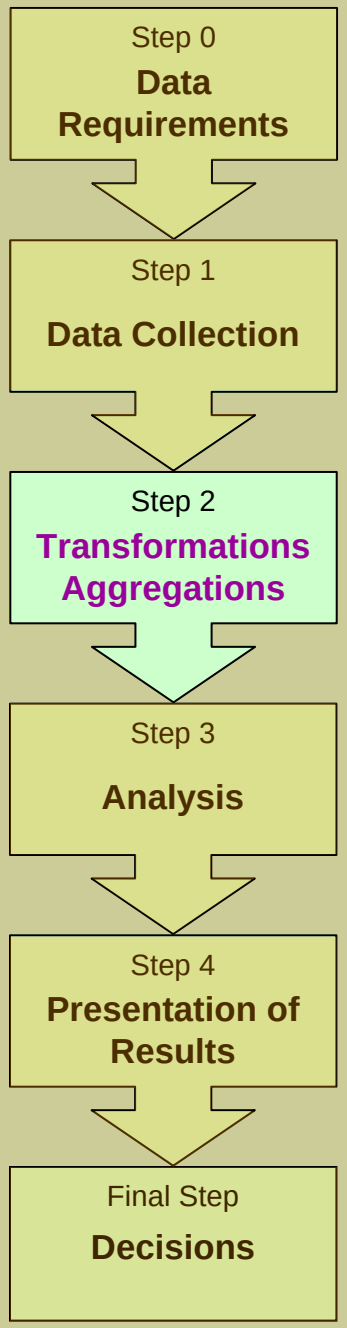
EDA: Summary

- Before data is analyzed, it is
  - Gathered
  - Cleaned
  - Integrated
- EDA Techniques used to
  - Explore the data
  - Detect missing values
  - Identify invalid values
  - Highlight outliers



EDA: Summary

- Use histograms, descriptive statistics and frequency tables
- For large data bases
  - Concepts same
  - More automated



Data Audits

- ASOP No. 23 does not require actuaries to audit data, but should understand the process
- **Main Idea:** compare the data intended for use to its original source, e.g., policy applications or notices of loss
- **Accuracy:** follow a sample of statistical back to source documents
- **Completeness:** follow a sample of source documents (records) to the final report

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

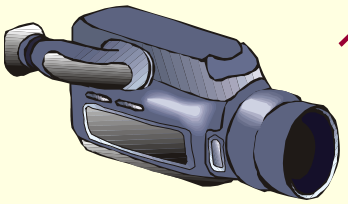
Step 4
Presentation of Results

Final Step
Decisions

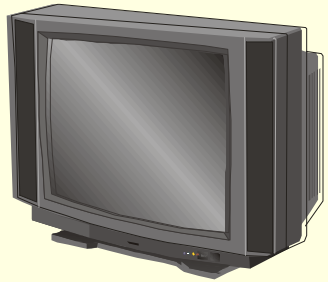
Analysis Quality

Models

Data



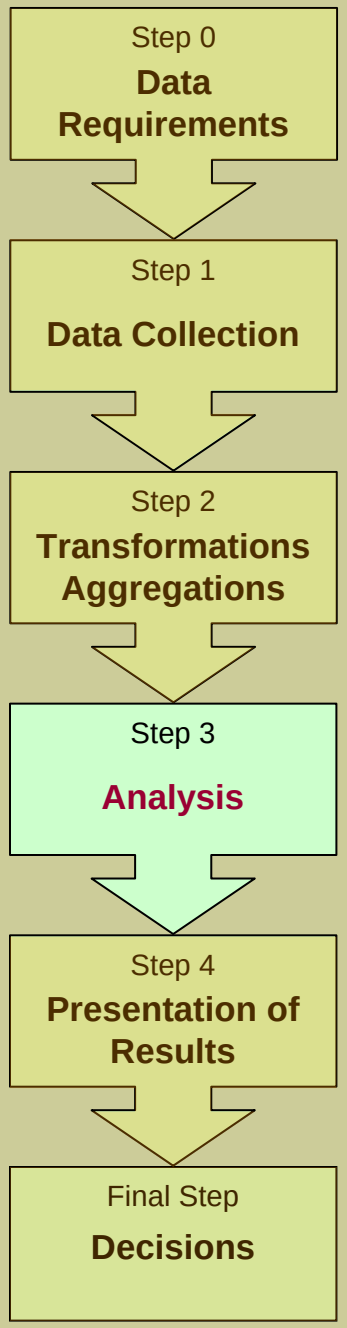
Results



On its way to results data can be:

- **Rejected**
  - **wrong Format**
- **Underutilized**
  - **wrong Model**
- **Distorted**
  - **wrong model Parameterization**

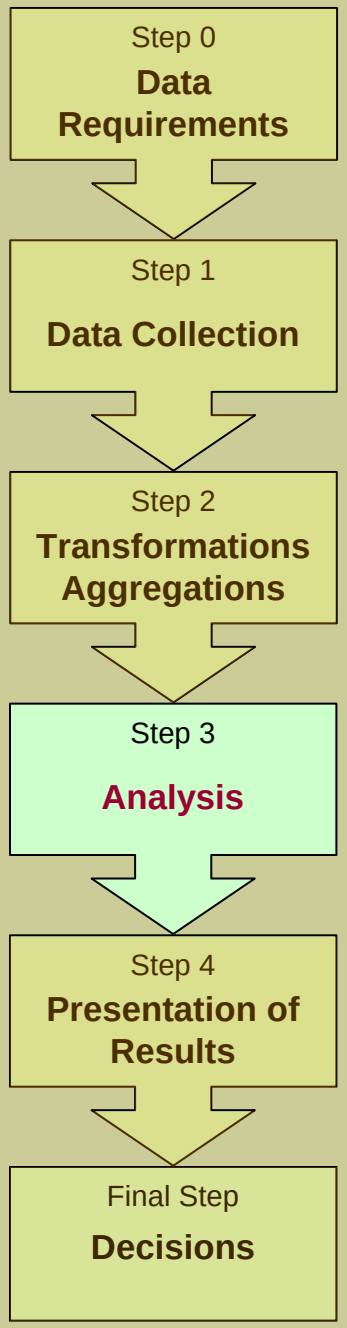
Analysis is a crucial component in the overall process quality



Model Quality

Section Agenda:

- Model design quality
- Implementation quality
- Testing and documentation



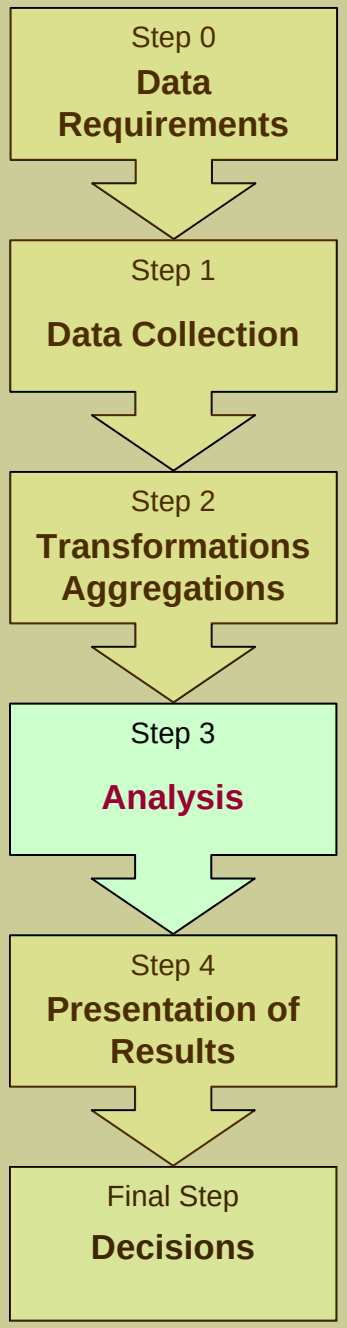
Model Quality

- Model Design quality
  - Model Selection and Validation
  - Parameters Estimation
  - Verification

Did I use the right model ?

Did I use the model right ?

  - Model Performance



Model Quality

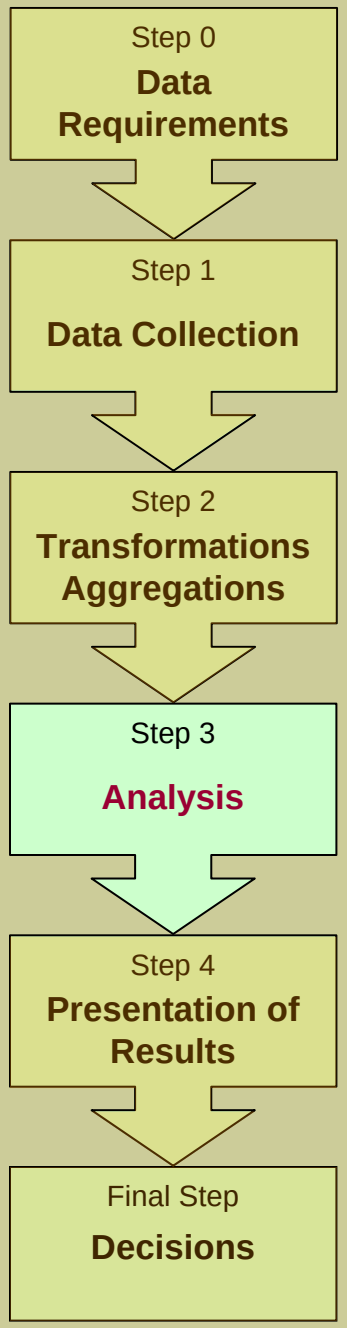
- Model Performance

Models predict observable events.

Outcomes can be compared to predictions leading to…

- Model Improvements
- Model Recalibration
- Model Rejection

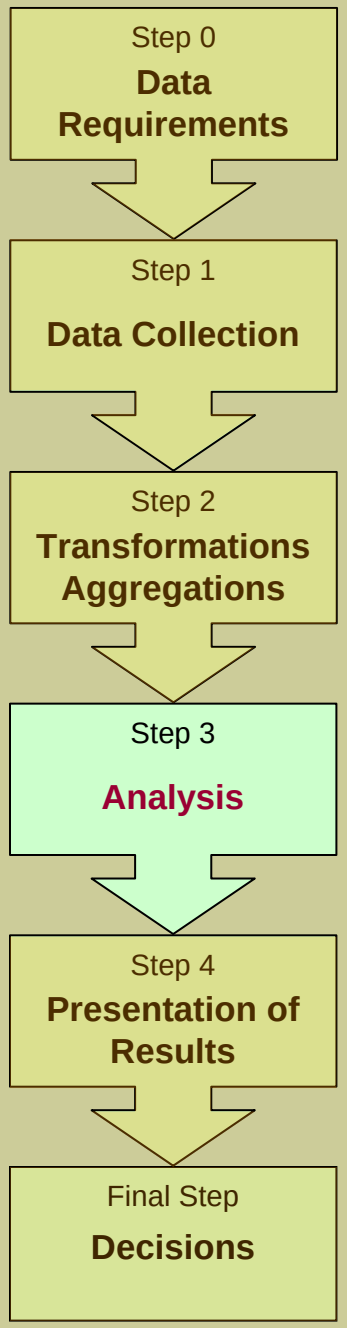
leading to… higher process quality.



Model Quality

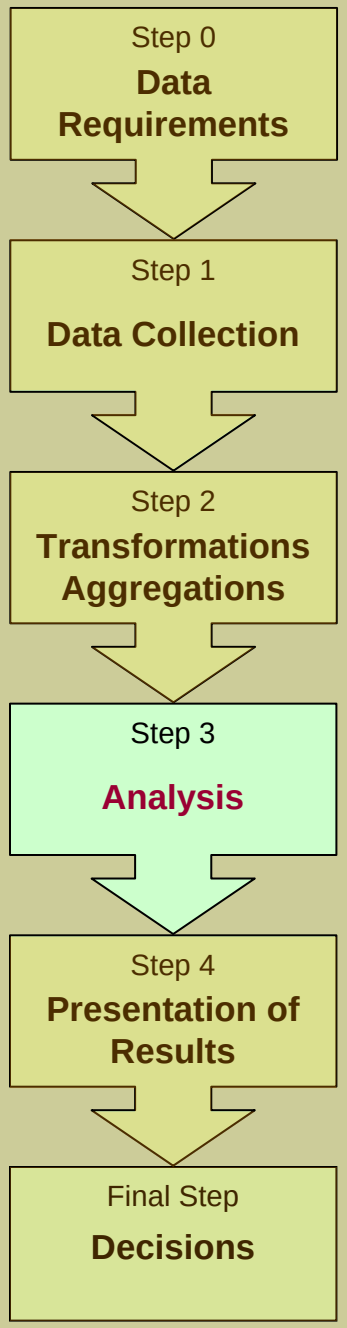
Section Agenda:

- Model Design quality
- Implementation quality
- Testing and Documentation



Model Quality

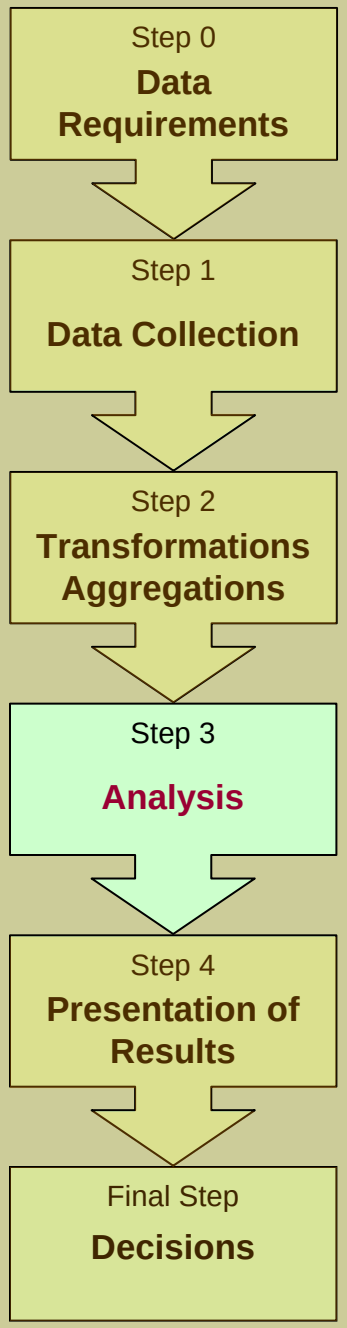
- Implementation quality
  - Programming languages: C++, VBA, SQL
    many books on good design patterns
  - Formulae in a Spreadsheet - also programming
    no books on good design patterns
  - Need good software design to simplify:
    - Usage
    - Testing
    - Modifications / Improvements
    - **Recovery** ← (side benefit)



Model Quality

Section Agenda:

- Model Design quality
- Implementation quality
- Testing and Documentation



Model Quality

- Testing and Documentation
  - Validation

black-box treatment: comparing results with correct ones…

  - Verification

inside-the-box treatment: checking formulae…

1. Should be **integral** part of development
2. Should be performed **by outsiders**
3. Should be **well-documented**

AGENDA

- Introduction
- Data Life Cycle
- **Data Management Best Practices**
- Conclusions

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

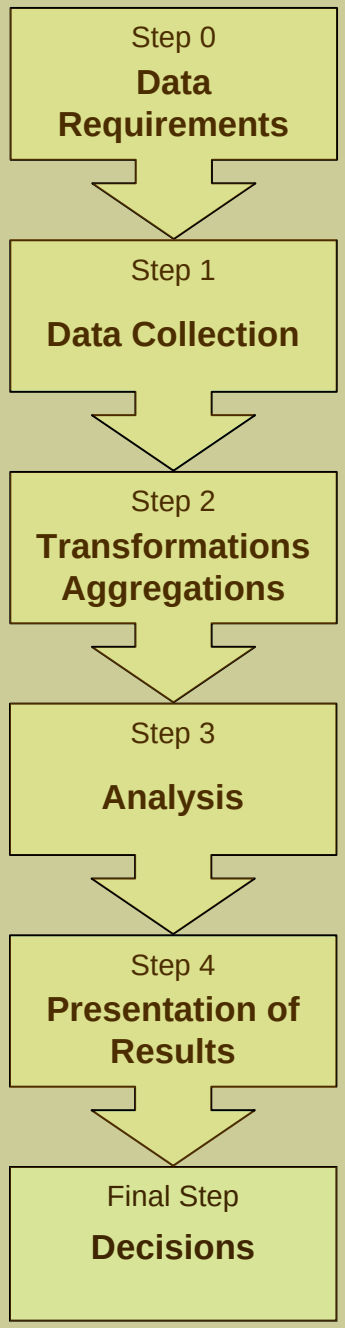
Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

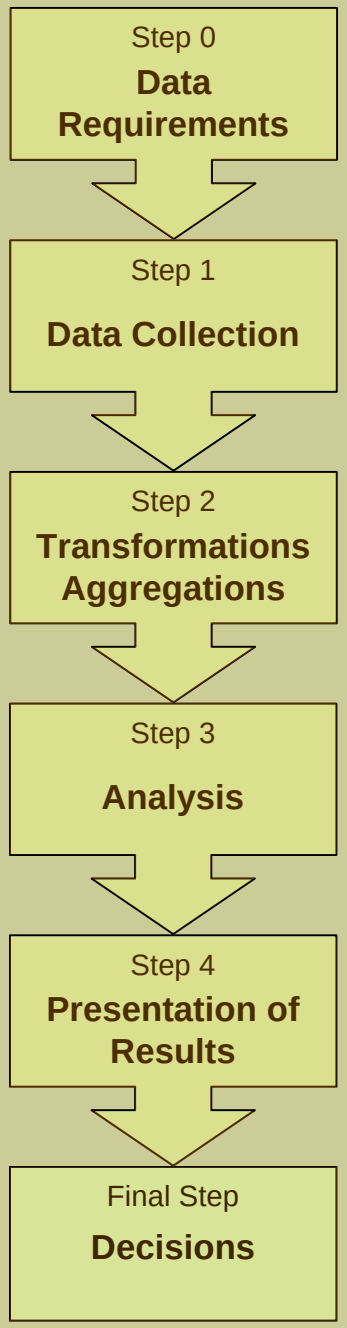
Actuarial Data Management

Bridge between data requirements, data collection, data transformation and aggregation, and data usage



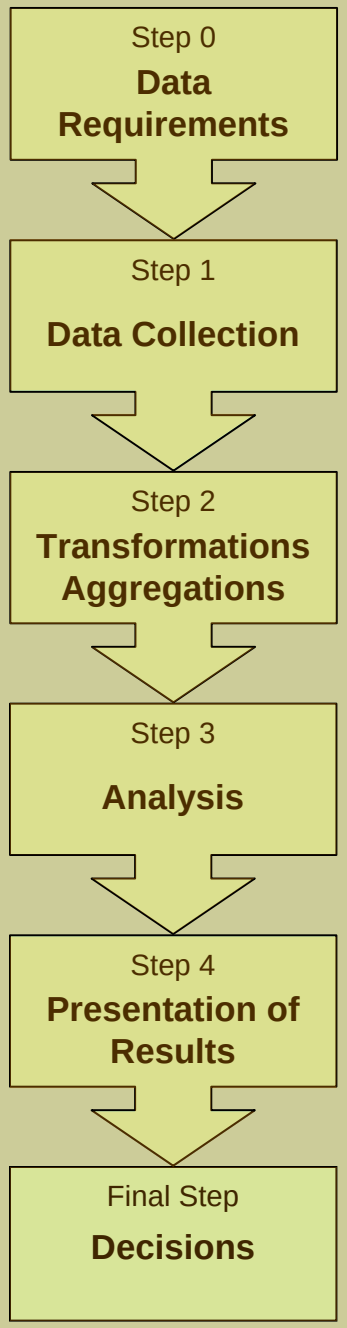
Critical Data Management Issues

- **Appropriateness** of the collected data elements for the related analyses
- **Quality** of the collected statistical experience for the related analyses



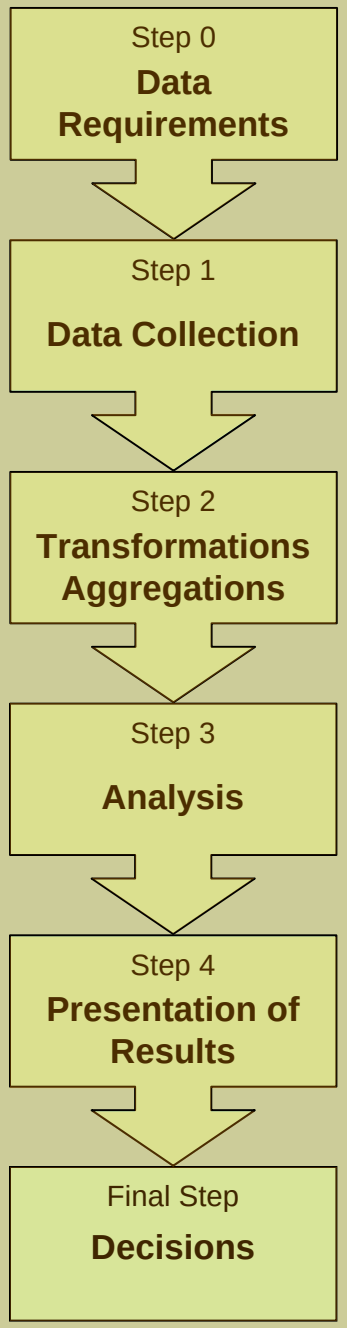
Data Management Best Practices & Guiding Principles

1. Data must be fit for the intended business use:
   - Even high quality data when repurposed may result in lessened data quality
2. Data should be obtained from the authoritative and appropriate source:
   - Data should flow from underlying business processes – example, expecting claim adjusters to create injury diagnoses
   - Know your data sources and their data quality and data management processes



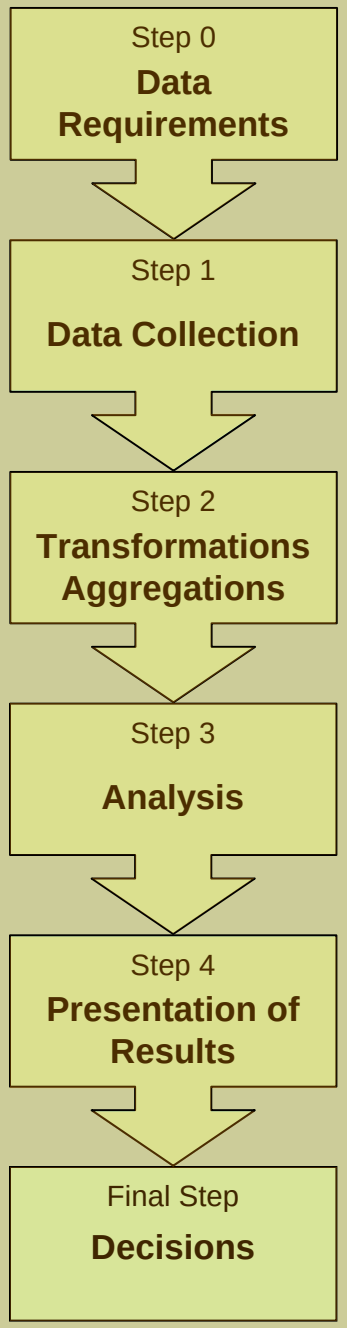
Data Management Best Practices & Guiding Principles

3. Common data elements must have a single documented definition and be supported by documented business rules:
   - B.I.: business intelligence, bodily injury, business interruption, ...
   - Incurred Loss: net as to deductible, net as to reinsurance, loss and expense, …
4. Metadata must be readily available to all authorized users of the data:



Data Management Best Practices & Guiding Principles

5. Data standards are key building blocks of DQ. Industry standards must be consulted and reviewed before a new data element is created:
   - Common Insurance Terminology (i.e., provision vs. reserve; what is a claim)
   - Coverage and Forms (i.e., motor vs. auto insurance)
   - Process Standards: Application Forms, Report of Injury or Claim, Licensing, etc.



Data Management Best Practices & Guiding Principles

- Solvency Standards – greatly impacting actuaries – Solvency II, RBC
- Data Exchange/Reporting Standards – external sources vs. internal data
- Data Quality Standards – industry DQ tools and report cards
- Data Element and Code List Definitions

Step 0
Data Requirements

Step 1
Data Collection

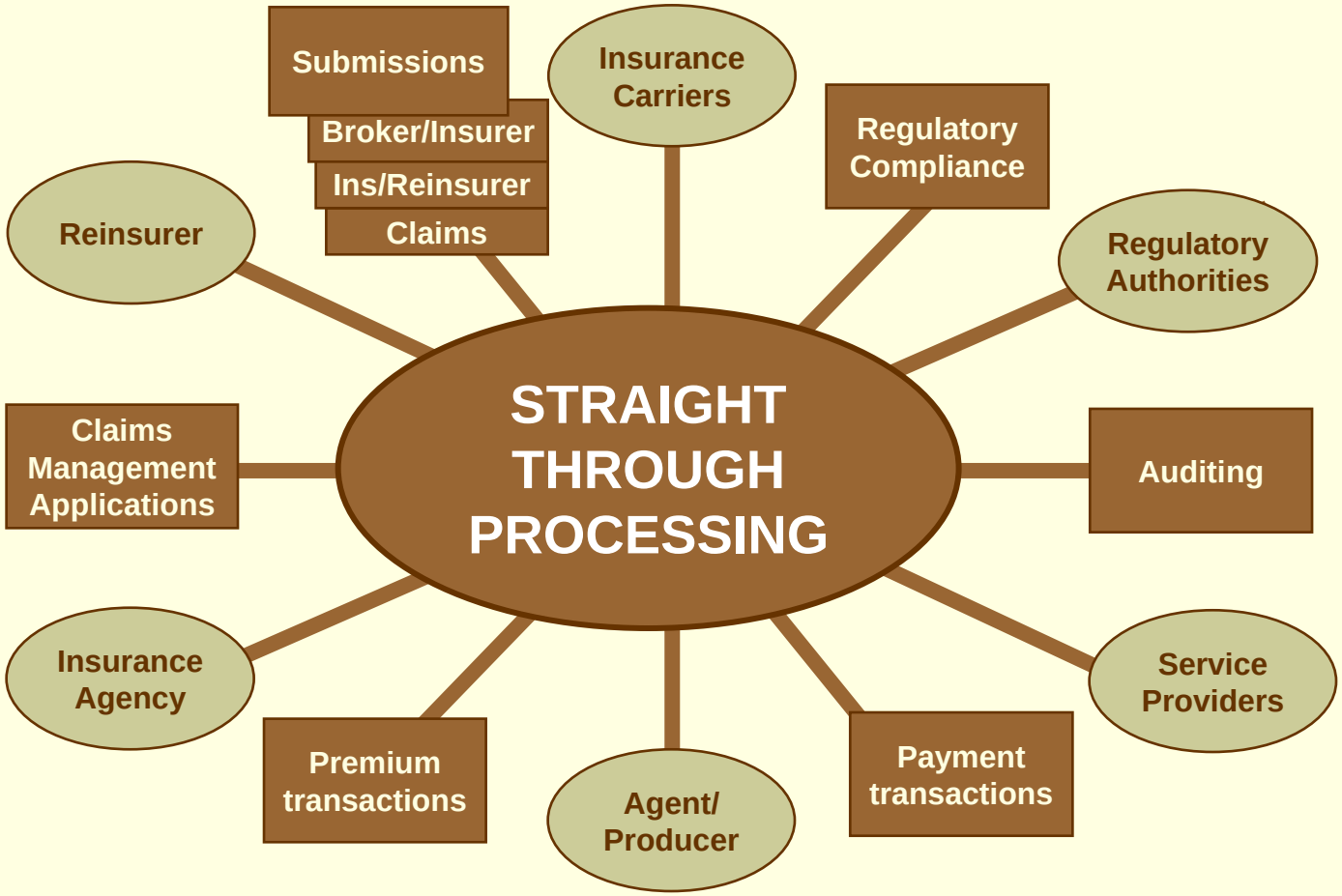
Step 2
Transformations Aggregations

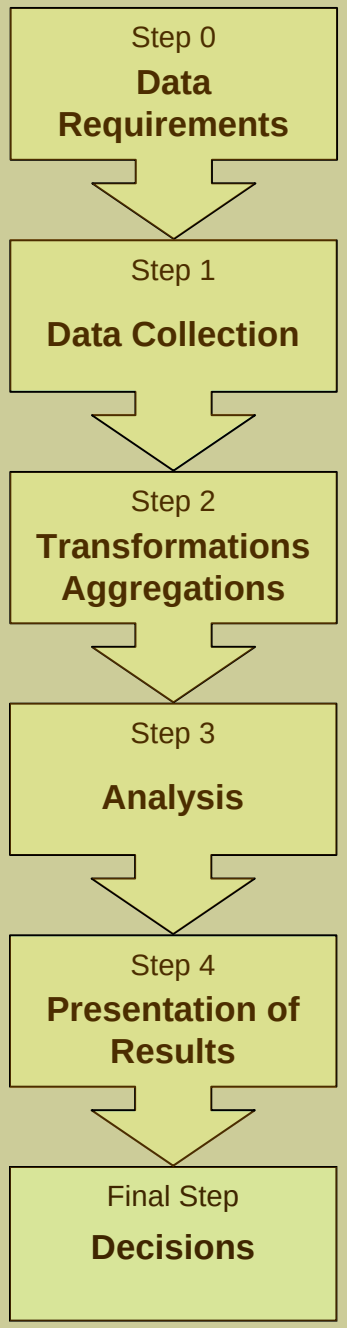
Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

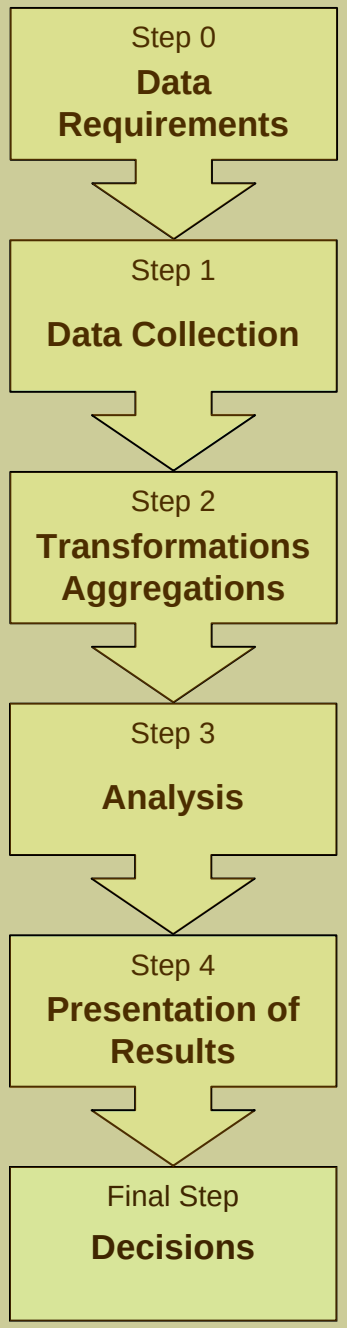
Benefits of Industry Data Standards





Data Management Best Practices & Guiding Principles

6. Data should have a steward responsible for
   - defining the data,
   - identifying and enforcing the business rules,
   - reconciling the data to the benchmark source,
   - assuring completeness, and
   - managing data quality.



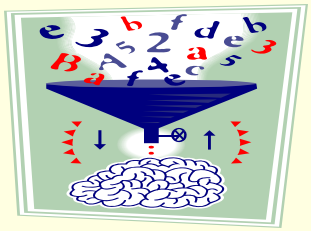
Data Management Best Practices & Guiding Principles

7. Data should be input only once and edited, validated, and corrected at the point of entry.
8. Data should be captured and stored as informational values, not codes.
9. Data must be readily available to all appropriate users and protected against inappropriate access and use.

AGENDA

- Introduction
- Data Life Cycle
- Data Management Best Practices
- **Conclusions**

PWC 2004 Study



“The key is to understand the impact data is having on your business and do something about it.”

“Data quality is at the core – if you improve your data you will directly impact your overall business results.”

Global Data Management Survey 2004, PriceWaterhouseCoopers

Conclusions

- Data Quality is a core issue affecting the quality and usefulness of the actuarial work products
- Data Quality is not just about how data is coded: the phrase “information quality” is coined to emphasize that processes impact the quality of the final product

Conclusions

- Ways to improve actuarial IQ discussed in the paper:
  - Applying Data Quality principles
  - Defining and using Metadata
  - Measuring data quality to track progress and awareness of quality audit
  - Using Exploratory Data Analysis to identify outliers and explore the structure of a dataset
  - Testing the quality of actuarial models
  - Clarifying actuarial presentations and reports
  - Employing IDMA's Data Management best practices

Conclusions

- Expanding actuaries' DQ perspective:
  - Data is a corporate asset that needs to be managed and actuaries can play a role
  - Data needs to be appropriate for all of its intended uses
  - Expand interpretation of data quality principles to support these broader perspectives

References

- **Actuarial Standard of Practice No. 23: Data Quality:** http://www.actuarialstandardsboard.org/pdf/asops/asop023_097.pdf
- **CAS DMIC Data Quality White Paper:** http://www.casact.org/pubs/forum/97wforum/97wf145.pdf
- **Insurance Data Management Association:** www.idma.org

Author, Author…

This presentation is a publication of CAS

Data Management and Information Educational Materials Working Party:

- Keith P. Allen
- Robert Neil Campbell, *Chairperson*
- Louise A. Francis
- David Dennis Hudson
- Gary W. Knoble
- Rudy A. Palenik
- Aleksey Popelyukhin Ph.D.
- Virginia R. Prevosto
- Lijuan Zhang

CAS Data Management Educational Materials Working Party Publications

- **Book reviews** of data management and data quality texts in the CAS *Actuarial Review* starting with the August 2006 edition
- These reviews are combined and compared in "**Survey** of Data Management and Data Quality Texts," CAS *Forum,* Winter 2007, www.casact.org

This presentation is based on our recently published **paper**:

- "**Actuarial IQ** (Information Quality)" published in the Winter 2008 edition of the CAS *Forum:* http://www.casact.org/pubs/forum/08wforum/