



Data Quality and Preparation For Model Building

By Dr. Arthur L. Dryver

Email: Dryver@gmail.com

URL: <http://as.nida.ac.th/~dryver>

National Institute of Development
Administration (NIDA)

August 11th 2006



First, Who Am I?

■ Work Experience:

- I am a lecturer at NIDA within the School of Applied Statistics. I started working at NIDA in October 2003.
- I worked in the US for 4 years within consulting. Companies I worked for:
 - **Experian, AnaBus, and PricewaterhouseCoopers (PwC)**
 - While working as a consultant I performed analyses on data from various industries.
 - **For several projects I had to work with multiple files with over a million records.**
- At Experian I had to create fraud detection models for clients.

■ Education

- The Pennsylvania State University
 - Ph.D. in Statistics 1999
- Rice University
 - BA in Mathematical Sciences/Statistics 1993



Outline

1. The Six Phases of Data Mining

2. Data Understanding Phase

1. With a focus on data quality
 1. What is data quality.
 2. How do we measure it.
 3. What can be done to increase data quality

3. Data Preparation Phase

1. With a focus on variable creation and data selection.

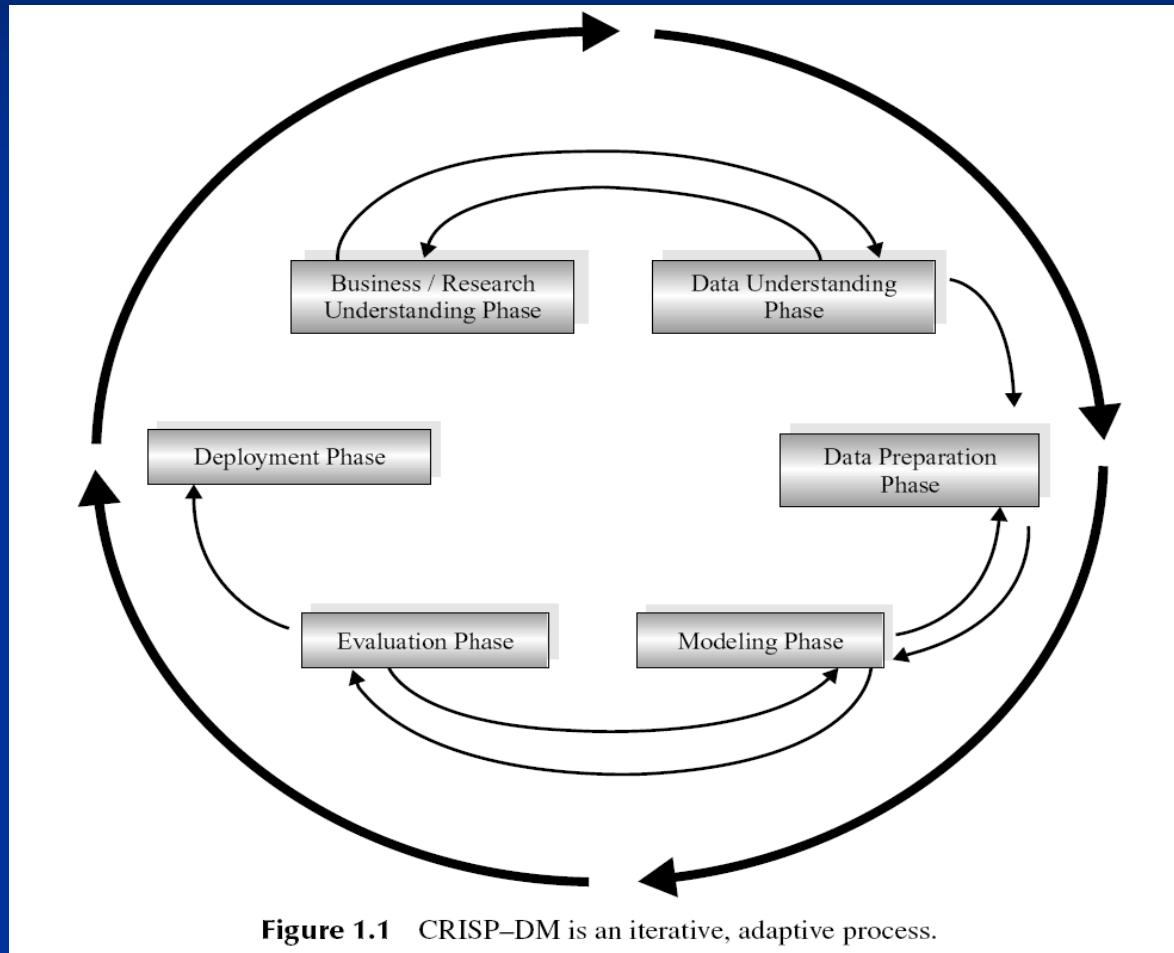


The Six Phases of Data Mining

The Cross-Industry Standard Process
for Data Mining (CRISP-DM)



The Six Phases of Data Mining [1]



[1] Excerpt: John Wiley & Sons - Discovering Knowledge in Data An Introduction to Data Mining by Daniel T. Larose



The Six Phases of Data Mining [1]

1. **Business understanding phase.** The first phase may also be termed the research understanding phase.
 1. Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole.
 2. Translate these goals and restrictions into the formulation of a data mining problem definition.
 3. Prepare a preliminary strategy for achieving these objectives.
2. **Data understanding phase**
 1. Collect the data.
 2. Use exploratory data analysis to familiarize yourself with the data and discover initial insights.
 3. Evaluate the quality of the data.
 4. If desired, select interesting subsets that may contain actionable patterns.



The Six Phases of Data Mining [1]

3. Data preparation phase

1. Prepare from the initial raw data the final data set that is to be used for all subsequent phases. This phase is very labor intensive.
2. Select the cases and variables you want to analyze and that are appropriate for your analysis.
3. Perform transformations on certain variables, if needed.
4. Clean the raw data so that it is ready for the modeling tools.

4. Modeling phase

1. Select and apply appropriate modeling techniques.
2. Calibrate model settings to optimize results.
3. Remember that often, several different techniques may be used for the same data mining problem.
4. If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique.



The Six Phases of Data Mining [1]

5. Evaluation phase

1. Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field.
2. Determine whether the model in fact achieves the objectives set for it in the first phase.
3. Establish whether some important facet of the business or research problem has not been accounted for sufficiently.
4. Come to a decision regarding use of the data mining results.

6. Deployment phase

1. Make use of the models created: Model creation does not signify the completion of a project.
2. Example of a simple deployment: Generate a report.
3. Example of a more complex deployment: Implement a parallel data mining process in another department.
4. For businesses, the customer often carries out the deployment based on your model.



The Six Phases of Data Mining [1]

- You can find out much more information about the CRISP-DM standard process at www.crisp-dm.org.
- This presentation will focus on parts within phase 2 and phase 3 of the six phases.
- An important fact to not forget is that CRISP-DM is an iterative adaptive process.



Data Understanding Phase

With a focus on data quality.



1: Collect the Data

1. Data mining is often performed on an existing data base. The data collection part is mainly talking to the appropriate people to obtain the data pertinent to business problem at hand.
 - This comment is from a consulting perspective.
2. The data you desire is obviously determined by the questions you desire to answer.
3. Data you have will in part determine the business issues you can address
 1. To answer certain questions if the data is not available you may need to collect additional data. The collection of additional data can be costly.
 2. Sometimes on a project it may be necessary to modify your business question according to the information at hand.
 - For example, revenue may be used as a proxy for profit. Revenue is clearly not profit, but if profit is not feasible to obtain, then revenue may be the next best measure.
4. **In this step it is very important to talk with people closest to the questions at hand to get the best insight into the data required to answer those questions.**



2: Use Exploratory Data Analysis To Familiarize Yourself With the Data and Discover Initial Insights

1. In this step it is very important to recall what the people closest to the questions at hand recommended to investigate.
 1. Especially when time is a concern, this will help you focus on possibly the most important variables and ensure you don't miss any obviously important variables due to rushing.
 2. From personal experience, if there are certain variables people involved in the project believe are relevant then at the end of the project questions will arise pertaining to those variables.
 1. It is very important to be able to at least comment on why a variable is not used in the final solution if another person believed the variable was relevant.
 2. This could cause doubt in the solution if a certain variable perceived to be important was not investigated. Possibly causing extensive additional work, possibly even resulting in a different solution.



2: Use Exploratory Data Analysis To Familiarize Yourself With the Data and Discover Initial Insights

2. Often I have worked with databases on a short timeline and with several files, hundreds of thousands of records each file, and hundreds of variables. In these circumstances variables mentioned by the client will be focused on.
 1. The minimum, maximum, mode, mean, median, 25th and 75th percentiles will be of most interest to get a quick understanding.
 - Again, additional time and thought will be spent on the key variables mentioned by the client.
 2. When dealing with such a large number of variables the modeling phase will often be used for the elimination of variables.



3: Evaluate the Quality of the Data

- There are many aspects to cover when thinking about the quality of the data. I will cover a few important ones.
 - A. Do you have the data needed to answer the questions you have?
 - B. Is the data accurate?
 - C. Is a large percent of data missing within the database?
 - D. Can you make the most use of your data?
 - Can you combine your databases to make the most use of all your data?
 - E. Do you have consistent labeling and naming conventions within the database?



A. Do you have the data needed to answer the questions you have?

1. A database is of little use if it is missing key elements to answer the questions at hand. Thus making the database in essence low quality.

Example:

- Company “X” is keeping a database for marketing purposes. They wish to use the database to learn about the success of their marketing campaigns to improve future campaigns. If their database does not contain information about items purchased as a result of their marketing campaigns this would make it almost impossible to learn from previous campaigns.

2. Sometimes worthless data exists.

- A National database with a field for nationality. You do not need a field for a million records with the same response for every record.



B. Is the data accurate?

- Believe it or not many databases keep a lot of inaccurate information.
- I know of one company paid a lot of money for advanced statistics done on data they knew was unreliable.
 - Don't spend a lot of money on advanced statistics if you know your data is inaccurate.
 - This creates a politically driven solution.
 - If the company likes the solution they ignore the fact the data is inaccurate, and if they don't like the solution they ignore the solution and mention the fact the data is inaccurate.
- Often we look at the minimum, maximum, least frequent, and most frequent observations to get a quick feel for reliability of the data.



C. Is a large percent of data missing within the database?

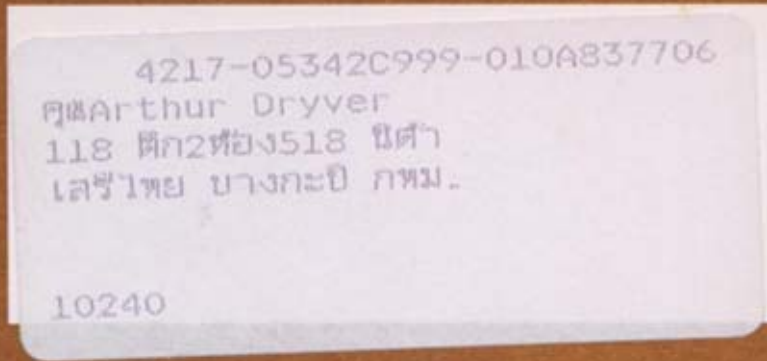
- Are you missing a lot of data?
- Are essential fields missing data?



D. Can you make the most use of your data?

1. Example, a fast food company “X” that delivers food, mailed out surveys to learn about their customers.
2. I received a survey. They have my information on file since I order food from them and probably have a lot of information already about my eating habits.

1: The front of the survey, has a sticker with my address for mailing it to me.



2: In the beginning of survey requesting identification information:

Name: XXXXX

Address XXXXXX

Phone: XXXXXXX

3: The survey is approximately 10 pages long. Questions range from gender, income, to food preferences. **Valuable information.**

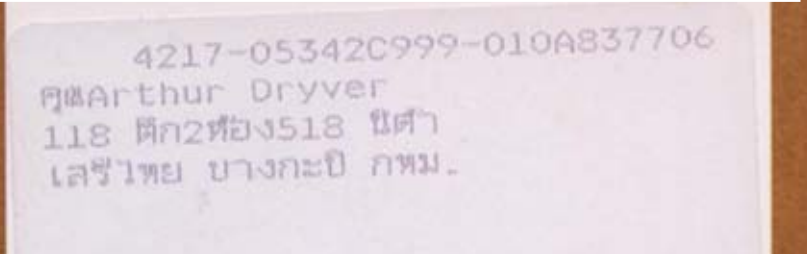
4: Think about a potential issue with the way in which the data is collected.
Take a few minutes.



D. Can you make the most use of your data?

- They have my information on file since I order food from and probably have a lot of information already about my eating habits.

The front of the survey, has a sticker with my address for mailing it to me.



1. The sticker was not a peel off for me to re-use, thus I had to write my information by hand on the survey.
2. As someone with a lot of data experience I know this is not good!

How do you merge the data from the survey back to the main database?

- Name: How do you merge Author Driver and Arthur Dryver? Data is often not entered correctly. Once at work they called me about spelling my name since it was spelt Author Driver on some documents.
- Phone number: What if I don't give my phone number or my phone number changed.

Gathering data can be costly, make the most of your data. Don't lose 10-30% due to poor vision. Place a unique identification number on the sticker for the respondent to write place in the survey, if they are the person with the name on the survey.



D. Can you make the most use of your data?

- A more common problem. Many companies have high churn, customers leaving for another company and many of their former customers later return.
 - Imagine cell phone service providers, such as DTAC, AIS, etc.
 - Last month you had service with company “X”, this month you use company “Y”, but you decided to switch back to “X”. Will “X” know you’re a former customer. Would they know if you didn’t pay your last months bill?
- Many companies have a customer on their database listed multiple times and can’t determine when it is the same person.
 - I have done a lot of work on databases and often been told that each record in the database refers to a unique person, only to later find the same person listed multiple times.



E. Do you have consistent labeling and naming conventions within the database?

- Many databases do not have a consistent labeling convention.
 - Imagine a database with province/city:
 - Bangkok
 - BKK
 - bkk (some software packages are case sensitive)
 - Krung Thep
 - กรุงเทพฯ
 - Terrible for data analysis. This can lead to difficulties in learning about Bangkok. What if you don't realize there are multiple labels for Bangkok? This can add hours and/or result in inaccurate statistics if you want to summarize by province/city.



4: If Desired, Select Interesting Subsets That May Contain Actionable Patterns

1. Sometimes you may wish to eliminate some data after exploratory data analysis if you feel it adds no value toward the questions at hand and will actually add difficulty to the modeling phase.



Data Preparation Phase

With a focus on variable creation and
data selection.



Variable Creation

Often we have to create new variables from the data we are working with.



Example 1

Payment History



Payment History: How can we use this information to understand risk? Take a few minutes.

- Two year payment history (24 months, present to 23 months ago)
 - 0=current, not late on payment
 - 1=30 days late
 - 2=60 days late
 - 3=90 days late
 - 4=120 days late, in default
 - This person's information has been given to a collection agency. The collection agency, is another company that will try to collect the money owed.

A snapshot of the data. Status0 is the present status. Status23 is the person's 23 months. You have all 24 months, just not shown.

Name	Status23	Status22	Status4	Status3	Status2	Status1	Status0
JINSMGK	3	4	4	4	4	4	4
WQOMEWTY	0	1	0	0	0	0	0
TSLOKOFY	0	0	0	0	0	0	0
YIBIJHB	2	3	0	0	0	0	0
DTHEE	0	0	0	1	2	0	0



Payment History: How can we use this information to understand risk? It is necessary to create new variables for modeling, cannot use as is. Take a few minutes to think.

■ **Two year payment history (24 months, present to 23 months ago)**

- 0=current, not late on payment
- 1=30 days late
- 2=60 days late
- 3=90 days late
- 4=120 days late, in default

A snapshot of the data. Status0 is the present status. Status23 is the person's 23 months. You have all 24 months, just not shown.

- This person's information has been given to a collection agency. The collection agency, is another company that will try to collect the money owed.

Name	Status23	Status22	Status4	Status3	Status2	Status1	Status0
JINSMGK	3	4	4	4	4	4	4
WQOMEWTY	0	1	0	0	0	0	0
TSLOKOFY	0	0	0	0	0	0	0
YIBIJHB	2	3	0	0	0	0	0
DTHEE	0	0	0	1	2	0	0



Payment History: There are a lot of possibilities for new variables.

1. Indicator variables of status0 (Present Status), Examples:
 1. Present status current or not.
 2. Present status 60 days or more.
 3. Present status in default or not.
2. Worst Ever. What is the worst (maximum) status over the past 24 months (Indicator variable).
3. Worst status past 12 months. What is the worst (maximum) status over the past 12 months (Indicator variable).

Name	Status23	Status22	Status4	Status3	Status2	Status1	Status0
JINSMGK	3	4	4	4	4	4	4
WQOMEWTY	0	1	0	0	0	0	0
TSLOKOFY	0	0	0	0	0	0	0
YIBIJHB	2	3	0	0	0	0	0
DTHEE	0	0	0	1	2	0	0



Payment History: There are a lot of possibilities for new variables.

4. Worst Ever (maximum) status is 60 days or more.
5. Worst Ever (maximum) status is 90 days or more.
6. Sum of status23 to status0.
 - Very bad. Why, what is wrong with this?
7. Average of status23 to status0.
 - Very bad. Why, what is wrong with this?

Think about why 6 and 7 are not a good measure of risk.

Name	Status23	Status22	Status4	Status3	Status2	Status1	Status0
JINSMGK	3	4	4	4	4	4	4
WQOMEWTY	0	1	0	0	0	0	0
TSLOKOFY	0	0	0	0	0	0	0
YIBIJHB	2	3	0	0	0	0	0
DTHEE	0	0	0	1	2	0	0



Payment History: There are a lot of possibilities for new variables.

6. Sum of status23 to status0.
 - Very bad. Why, what is wrong with this?
7. Average of status23 to status0.
 - Very bad. Why, what is wrong with this?

Think about why 6 and 7 are not a good measure of risk. It is possible with this measure of risk for a person in default to have a lower average score than someone who is presently current and not late. Thus these two variables can yield a very inaccurate picture of true risk.

Name	Status23	Status22	Status4	Status3	Status2	Status1	Status0
JINSMGK	3	4	4	4	4	4	4
WQOMEWTY	0	1	0	0	0	0	0
TSLOKOFY	0	0	0	0	0	0	0
YIBIJHB	2	3	0	0	0	0	0
DTHEE	0	0	0	1	2	0	0



Payment History: There are a lot of possibilities for new variables.

6. Sum of status₂₃ to status₀.
 - Very bad. Why, what is wrong with this?
7. Average of status₂₃ to status₀.
 - Very bad. Why, what is wrong with this?

Below is an example of two people's payment history. From 24 months ago (23) to the present (0). The first person is presently in default and are written off as bad debt. The second person is last often but is presently current. The second person is better than first person. Looking at the sum or mean would give the opposite viewpoint and be misleading.

23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	sum	mean
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	4	10	0.42
0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	0	33	1.38



Example 2

Transaction Log Data (T-Log Data)



T-Log Data: How can we use this information to understand about the different configuration?

- Comparing different types of checkout counter styles and cash registers using transaction log data (T-log) in terms of speed.
- Partial T-Log Data:
 - Configuration of checkout counter.
 - There are 4 types. 2 Different shapes and 2 different cash register types.
 - nitems=number of items purchased during transaction
 - tender=0 if cash is used,1 if credit is used
 - massist=1 if manager assist 0 otherwise
 - timer1=time first item is scanned
 - timer2=time last item is scanned
 - timer3=time transaction completed

A snapshot of the data.

Configuration	Nitems	tender	massist	timer1	timer2	timer3
1	2	0	0	0	23.45	32.87
1	7	0	0	170.85	217.18	227.62
1	8	0	0	242.33	303.92	314.35
1	5	0	0	330.78	368.57	378.41
1	10	0	0	576.96	647.7	657.4



T-Log Data: It is necessary to create new variables for modeling, cannot use the data as is. Starting with creating the dependent variable.

- We would want to do a general linear model to investigate speed in terms of configuration for the different shapes and register types.
- Partial T-Log Data:
 - Configuration of checkout counter.
 - There are 4 types. 2 Different shapes and 2 different cash register types.
 - nitems=number of items purchased during transaction
 - tender=0 if cash is used,1 if credit is used
 - massist=1 if manager assist 0 otherwise
 - timer1=time first item is scanned
 - timer2=time last item is scanned
 - timer3=time transaction completed

A snapshot of the data.

Configuration	Nitems	tender	massist	timer1	timer2	timer3
1	2	0	0	0	23.45	32.87
1	7	0	0	170.85	217.18	227.62
1	8	0	0	242.33	303.92	314.35
1	5	0	0	330.78	368.57	378.41
1	10	0	0	576.96	647.7	657.4



T-Log Data: It is necessary to create new variables for modeling, cannot use the data as is.

- We would want to do a general linear model to investigate speed in terms of configuration for the different shapes and register types.
- An estimate for the time of a transaction could be a new variable equal to $\text{timer3} - \text{timer1}$.
- What about configuration. Really we would desire to variables, one variable for the shape of the counter and another variable for the register type.
- What else might be of interest?
 - Take a few minutes to think about.

A snapshot of the data.

Configuration	Nitems	tender	massist	timer1	timer2	timer3
1	2	0	0	0	23.45	32.87
1	7	0	0	170.85	217.18	227.62
1	8	0	0	242.33	303.92	314.35
1	5	0	0	330.78	368.57	378.41
1	10	0	0	576.96	647.7	657.4



T-Log Data: It is necessary to create new variables for modeling, cannot use the data as is.

- We would want to do a general linear model to investigate speed in terms of configuration for the different shapes and register types.
- We might want to understand configuration when the store is busy. How do we know when a store is busy:
 - When there is very little time between timer1 at time point i and timer3 at time point $(i-1)$. “Very little” time is partially subjective.
 - This is the time between the first item scanned and the completion of the last transaction.
 - A small amount time would indicate that there is a line and thus the store is busy.

Configuration	Nitems	tender	massist	timer1	timer2	timer3
1	2	0	0	0	23.45	32.87
1	7	0	0	170.85	217.18	227.62
1	8	0	0	242.33	303.92	314.35
1	5	0	0	330.78	368.57	378.41
1	10	0	0	576.96	647.7	657.4



Data Preparation:

The importance of working with the correct data (Data Selection):

An ATM example

In addition, this is an example of how important it is to check your work and the work of those working for you.



- Your boss has asked you to determine the average withdrawal amount and number of withdrawals from an ATM your Bank owns at the Mall Bangkapi for September 1st before 3:00PM today. Naturally, you ask your statistician for the information from the database. At 2:50PM the following is what the statistician returns. Your Boss calls at 2:55PM for the results. What is the mean and number of withdrawals?

<i>September 1st Withdrawals</i>	
Mean	\$ 3,204.02
Standard Error	\$ 179.88
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 5,688.16
Sample Variance	32,355,148.36
Kurtosis	4.42
Skewness	(1.74)
Range	\$ 29,997.33
Minimum	\$ (19,997.33)
Maximum	\$ 10,000.00
Sum	\$ 3,204,022.41
Count	1000



- Your answer should be “Sorry, but we are still working on it.”
A withdrawal is the taking out of money from an ATM. It can only be recorded as a positive or a negative value but not as both. There must be deposits included in the data. You mention this to the statistician and tell him/her to fix this.

<i>September 1st Withdrawals</i>	
Mean	\$ 3,204.02
Standard Error	\$ 179.88
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 5,688.16
Sample Variance	32,355,148.36
Kurtosis	4.42
Skewness	(1.74)
Range	\$ 29,997.33
Minimum	\$ (19,997.33)
Maximum	\$ 10,000.00
Sum	\$ 3,204,022.41
Count	1000



Positive and Negative Numbers, doesn't make sense.



- The statistician returns and says it has been taken care of. Here is the updated work. What two things tell you that he has done poor work again?

Old

<i>September 1st Withdrawals</i>	
Mean	\$ 3,204.02
Standard Error	\$ 179.88
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 5,688.16
Sample Variance	32,355,148.36
Kurtosis	4.42
Skewness	(1.74)
Range	\$ 29,997.33
Minimum	\$ (19,997.33)
Maximum	\$ 10,000.00
Sum	\$ 3,204,022.41
Count	1000

Updated

<i>September 1st Withdrawals</i>	
Mean	\$ 4,179.50
Standard Error	\$ 106.76
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 3,376.08
Sample Variance	11,397,927.68
Kurtosis	(0.73)
Skewness	0.59
Range	\$ 10,000.00
Minimum	\$ -
Maximum	\$ 10,000.00
Sum	\$ 4,179,500.00
Count	1000



Old

<i>September 1st Withdrawals</i>	
Mean	\$ 3,204.02
Standard Error	\$ 179.88
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 5,688.16
Sample Variance	32,355,148.36
Kurtosis	4.42
Skewness	(1.74)
Range	\$ 29,997.33
Minimum	\$ (19,997.33)
Maximum	\$ 10,000.00
Sum	\$ 3,204,022.41
Count	1000

Updated

<i>September 1st Withdrawals</i>	
Mean	\$ 4,179.50
Standard Error	\$ 106.76
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 3,376.08
Sample Variance	11,397,927.68
Kurtosis	(0.73)
Skewness	0.59
Range	\$ 10,000.00
Minimum	\$ -
Maximum	\$ 10,000.00
Sum	\$ 4,179,500.00
Count	1000

The Count should have changed if deposits were removed? Know your numbers!!!

Why is the minimum withdrawal \$0.00? What was withdrawn?



Old

September 1st Withdrawals	
Mean	\$ 3,204.02
Standard Error	\$ 179.88
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 5,688.16
Sample Variance	32,355,148.36
Kurtosis	4.42
Skewness	(1.74)
Range	\$ 29,997.33
Minimum	\$ (19,997.33)
Maximum	\$ 10,000.00
Sum	\$ 3,204,022.41
Count	1000

Correct Version

September 1st Withdrawals	
Mean	\$ 5,204.86
Standard Error	\$ 105.01
Median	\$ 5,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 2,975.60
Sample Variance	8,854,178.38
Kurtosis	(0.94)
Skewness	0.69
Range	\$ 8,500.00
Minimum	\$ 1,500.00
Maximum	\$ 10,000.00
Sum	\$ 4,179,500.00
Count	803



The "Correct Version". How do I know? Honestly, you would probably still be concerned given the mistakes that were and not fully sure. Looking at the minimum, a little high, but reasonable. The maximum looks possible as well. At least nothing is obviously wrong anymore. You might ask for a printout of the 1st 30 or so records in the file to look over the raw data, this also often helps.



September 1 st Withdrawals	
\$	5,000.00
\$	3,000.00
\$	10,000.00
\$	1,500.00
\$	10,000.00
\$	3,000.00
\$	3,000.00
\$	-
\$	3,000.00
\$	3,000.00
\$	5,000.00
\$	1,500.00
\$	3,000.00
\$	1,500.00
\$	1,500.00
\$	3,000.00
\$	5,000.00
\$	10,000.00
\$	10,000.00
\$	10,000.00
\$	(14,665.53)
\$	-
\$	3,000.00
\$	-
\$	5,000.00
\$	-
\$	5,000.00
\$	10,000.00
\$	-
\$	(14,166.63)

Raw data has no positive numbers below \$1,500 and no positive numbers above \$10,000. At least in the 1st 30 records. This adds confidence that the results are correct.

September 1 st Withdrawals	
Mean	\$ 5,204.86
Standard Error	\$ 105.01
Median	\$ 5,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 2,975.60
Sample Variance	8,854,178.38
Kurtosis	(0.94)
Skewness	0.69
Range	\$ 8,500.00
Minimum	\$ 1,500.00
Maximum	\$ 10,000.00
Sum	\$ 4,179,500.00
Count	803



Old

September 1st Withdrawals	
Mean	\$ 3,204.02
Standard Error	\$ 179.88
Median	\$ 3,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 5,688.16
Sample Variance	32,355,148.36
Kurtosis	4.42
Skewness	(1.74)
Range	\$ 29,997.33
Minimum	\$ (19,997.33)
Maximum	\$ 10,000.00
Sum	\$ 3,204,022.41
Count	1000

The true number of withdrawals is 803 as opposed to 1,000.
($1-803/1000 = .197$. 803 is 19.7% less than 1,000) **A huge difference!!!**

Correct Version

September 1st Withdrawals	
Mean	\$ 5,204.86
Standard Error	\$ 105.01
Median	\$ 5,000.00
Mode	\$ 5,000.00
Standard Deviation	\$ 2,975.60
Sample Variance	8,854,178.38
Kurtosis	(0.94)
Skewness	0.69
Range	\$ 8,500.00
Minimum	\$ 1,500.00
Maximum	\$ 10,000.00
Sum	\$ 4,179,500.00
Count	803

The true mean is \$5,204.86 as opposed \$3,204.02.
($5204.86/3204.02 - 1 = .624$) Thus \$5,204.86 is 62.4% larger than \$3,204.02. **A huge difference!!!**



Comments on ATM Example

- What do you think of the statistician working for you?
 - Not good. Almost correct is ok on class exams, but not when you are working. It must be correct. Small mistakes can lead to very different answers. This was one example.
 - Mistakes like this happen but the statistician should take time to check his work and think about it before handing you the results.
 - Just as you must think about the work before handing it to your boss.
 - People loose their job if they are careless like this statistician.
 - Trust is important, unfortunately you can't trust this statistician's work.



Comments on ATM Example

- Why is this important to you as a manager?
 - You as a manager are responsible for the work of the people you manage. Especially when you report/present the work they produce.
 - Yes, you can blame the statistician for his carelessness, but if you can not check the work of those that work for you, soon you will be considered a poor manager and will have a problem. Especially if your boss finds mistakes, he will question why couldn't you find the mistakes yourself?
 - When you buy something you count your change because you don't want to lose money. Not checking the work is like not counting your change, only if you trust the person, and even then simple checks should be done.



Data Selection: Outliers

Handling of outliers (observations with extremely different values than most other observations):



1: It is Highly Desirable To Create A Model That Can Give A Prediction For The Most Individuals

- In some situations you may wish to create a model that can predictive for most to all individuals that are in the population.
 - Outliers can cause a model to predict poorly for most observations to obtain better predictions on a few observations, the outliers.
 - It is dangerous to predict results for individuals with values on independent variables outside the range used in the model creation.
 - In this situation “capping” may be appropriate. Best explained with an example:
 - Income. If the 99% percentile for income is 150,000 baht/month, everyone with an income above 150,000 will have their income “capped” at 150,000 baht/month. Thus the new variable would have a maximum of 150,000 baht/month.
 - I have used capping for logistic regression, and would be considerably more cautious about using it in other situations. It could be bad to use otherwise.



2: Typical Situation

- In many situations you wish to understand the relationship between the dependent variable and independent variables or merely wish to predict/estimate for the general case.
 - Outliers can cause a model to predict poorly for most observations to obtain better predictions on a few observations, the outliers.
 - In this situation in data mining most of the time the outliers are deleted. In data mining of a database there is often more than enough data and deleting data will be the easiest/quickest way to obtain a basic understanding.
 - Again, it is dangerous to predict results for individuals with values on independent variables outside the range used in the model creation. Thus the model you created would not be suitable for ranges of “X” that were not used in the model creation.



Data Transformations

Reciprocal, Square, Square Root, Log, ...



Data Transformations

- When handling a databases with a lot of variables, in the hundreds, sometimes multiple transformations will be tried and a “brute force” method will be used. Example for a data analysis project with a dependent variable: continuous, dichotomous or ordinal.
 - You might transform every independent variable by Reciprocal, Square, Square Root, Log and then select a variable transformation or untransformed variable by which is most highly correlated with the dependent variable. A “brute force” method, sometimes used.



Conclusion

- Unfortunately in data mining and in any data analysis project there is no simple answer or formula to use.
- To have a successful data mining project you must:
 1. Listen carefully to all people closely involved in the project and those most knowledgeable about the data.
 2. Think every step through carefully. It can be very costly to rush and make mistakes.
 - Always remember, slow and correct is infinitely better than wrong and fast. Wrong and fast actually is negative value as it can yield misleading solutions.
 3. Finally, remember G.I.G.O. (garbage in garbage out), there is only so much you can do on an analytical project with bad data.