Master of Technology in Enterprise Business Analytics

Data Analytics Process:

Goal Setting and Data Preparation

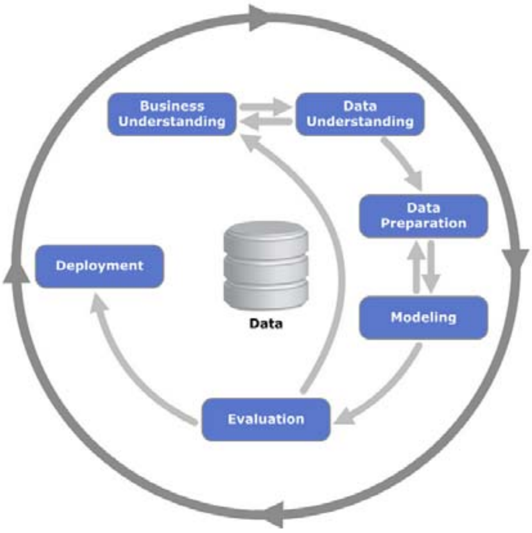
Dr. Barry Shepherd
Institute of Systems Science
National University of Singapore
E-mail: barryshepherd@nus.edu.sg

Data Analytics Process - Agenda

- Data Analytics Methodologies
- Goal Setting & Planning
  - Mini-workshop
- Plan Execution
  - Data Preparation
  - Model Building Process
  - Workshop/Assignment

The Data Analytics Process

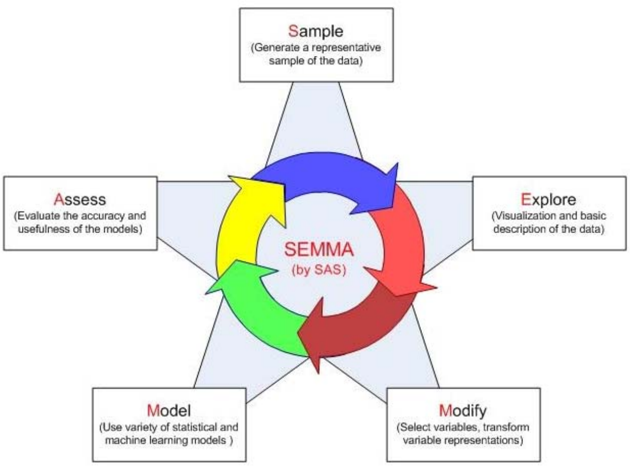
Many methodologies exist – mostly similar!



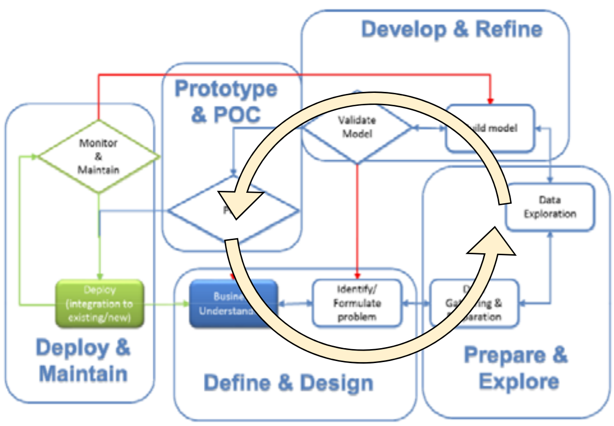
Source: Forrester Research, Inc.

We follow (mostly) the Cross Industry Standard Process for Data Mining (CRISP-DM).
(CRISP-DM was conceived in late 1996 by collaboration between vendors and end-user orgs, including SPSS, Daimler-Benz, NCR)

The Data Analytics Process



SEMMA, SAS



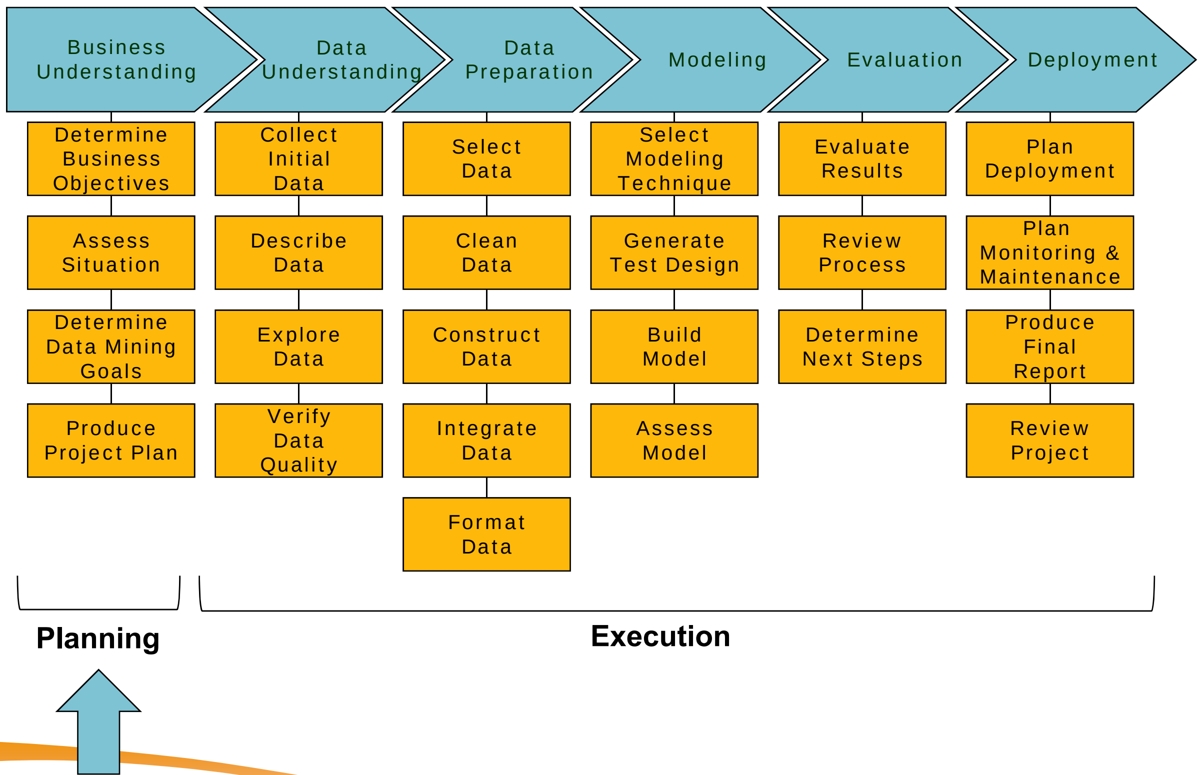
ISS, Catherine

Data Analytics Methodologies

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

Methodology	2014 poll	2007 poll
CRISP-DM (86)	43%	42%
My own (55)	27.5%	19%
SEMMA (17)	8.5%	13%
Other, not domain-specific (16)	8%	4%
KDD Process (15)	7.5%	7.3%
My organizations' (7)	3.5%	5.3%
A domain-specific methodology (4)	2%	4.7%
None (0)	0%	4.7%

CRISP-DM in Detail



Setting Business Goals

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

Determine Business Objectives

Assess Situation

Determine Data Mining Goals

Produce Project Plan

- Usually a ***two way*** process between the chief data scientist(s) and the business domain experts
  - The Data Scientist often needs ***some*** domain knowledge for this conversation to succeed

- Business Goal Guidelines
  - Use only business terms – make no mention of analytics methods!
  - There must be an actionable outcome
  - Must be able to measure success (quantifiable metrics)

Setting Business Goals

- Possible examples are …
  - Improve the response rate for a direct marketing campaign
  - Increase the average order size
  - Determine what drives customer acquisition
  - Forecast the size of the customer base in the future
  - Retain profitable customers
  - Recommend the next, best product for existing customers
  - Choose the right message for the right groups of customers

By how much?
Need clear success criteria

Assessing the Situation

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

Determine Business Objectives

Assess Situation

Determine Data Mining Goals

Produce Project Plan

- What is the background to the project?
- What analytics are already being done? (don't reinvent)
- What resources are available?
  - Data Sources (what is available? who owns it?, what is required?)
  - Software & Tools (any tools that must be used?)
  - Hardware (enough for data size?)
  - Knowledge sources (availability of domain experts)
- What are the constraints?
- What are the risks? contingency plan?

Setting Analytics Goals

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

Determine Business Objectives

Assess Situation

Determine Data Mining Goals

Produce Project Plan

- A business goal states objectives in business terminology
  - E.g. *"Increase catalog sales to existing customers."*
- A data mining goal states project objectives in technical terms.
  - E.g. *"Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city, etc.), and the price of the item."*

Requires an understanding of when and why to use the many different analytics techniques:

- When to use predictive modeling?
- When to use association mining?
- When to use clustering?
- etc…

Selecting an Analytics Approach

Does the problem map to a generic problem type?

- Are there suspected correlations, relationships ? — *Exploration/Visualisation*
- Is there something that could be useful to predict? — *Predictive Modelling*
- Do you hope to find things that happen (close) together? — *Association Finding*
- Do you want to compare the current situation with past situations? — *Memory-based*
- Do you hope/expect to find groupings/clusters? — *Statistical Clustering*
- Are there exceptional cases that need investigation? — *Outlier detection*
- None of the above – just find me some insights! — *Visualisation & Exploration*

Problem Type → **Analytics Approach** → **Analytics Techniques**

Setting Analytics Goals

- There may be many analytics approaches that meet a business goal
- ***Example:*** Company X wishes to increase sales to existing customers. How can we use analytics?

(1) Examine **past purchase data**, identify big spenders and target them with promotions.

(2) Examine customer profiles (demographics, interests etc.) – identify the big spenders, then target low spenders who ***"look like"*** the big spenders

(3) Examine the records of **past marketing campaigns,** combine this with **customer profile data** and **past purchase data** to build a ***response model*** to predict which customers will respond best to new campaigns

Setting Analytics Goals

- ***Example***: You are the marketing VP for a bank and your primary business objective is to retain current customers who are at risk of moving to a competitor.

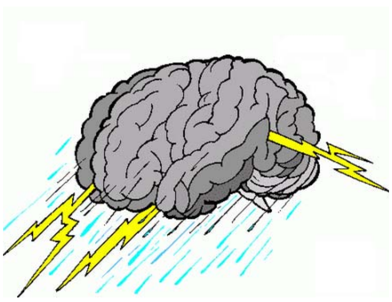
- Possible Approaches
  - Identify ***likely churners*** then offer them incentives to stay.
    To do this get customer profile data and account usage data for both loyal customers and churners. Use this to build a churn prediction model.
  - Identify ***the issues causing customers to churn*** – then fix these issues!
    What data is required for this?

Setting Analytics Goals

- ***Example:*** You are the marketing VP for a bank and you wishes to identify the top issues causing customers to churn so that they can be fixed
- Possible Approaches:

Searching for the issues by an
un-directed analysis of the data can be hard

Often better to brainstorm possible issues
and then use analytics to verify and rank them



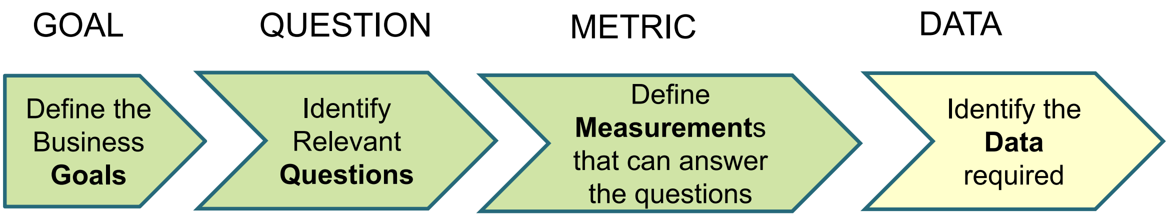
Channel influence on churn - How does the interaction channel (e.g. ATM, branch, or Web) affect loyalty & churn?

ATM pricing association - Will lower ATM fees significantly reduce the number of high-value customers who leave?

ATM pricing association with customer segments - Will lower fees affect only one particular customer segment?

Setting Analytics Goals

- The GQM Method is a good framework for brainstorming
  - Originally developed to help an organisation identify appropriate software metrics*



- Try to identify all of the known business issues related to your strategic objective to ensure that your data mining project is as business-focused as possible.

[1] Victor Basili, *Software Modeling and Measurement: The Goal/Question/Metric Paradigm*, CS-TR-2956, University of Maryland, 1992

Identifying Data Requirements

- Questions to Answer:
  - What data is available?
  - What must the data contain?
  - What would be useful? (whether available or not) – **innovate!**
  - What is the right level of granularity?
  - What volume of data is needed?
  - How much history is required?
    how far back in time should the data go?
- What data is required for comparison?
  - What is currently being done?
    - E.g. what is the existing churn rate, response rate, failure rate?
  - Obtain a control group ~ data describing the status quo
    - E.g. what happened to patients who did not receive the treatment?
    - E.g. what did customers buy who did not see the ad?

?

Identify any Data Gap

- Consolidate all of your data requirements
- Determine what (if any) essential data is missing
- How to bridge the gap?
  - Put in place mechanisms to start collecting the missing data (delay the analytics)
  - Get the data from elsewhere (e.g. 3rd party, the web)
  - Innovate to obtain missing data or data you think may be useful

Scenario: Public Transport Optimisation

- Public transport is a hot topic in Singapore, increasing population is driving the need for optimisation and innovation
  - One problem is **Bus Overcrowding** ~ what are the root causes?
  - Another problem is how to ensure **Bus Lanes** are effective. What are the characteristics of a successful bus lane?

Photo: ST

Example: Bus Lane Effectiveness

- **Business Problem:**
  - Will my planned bus lane(s) be effective?
  - If not effective then why not? Can it be fixed & how?
- **Business Goal:**
  - Make a go/no-go decision on a planned new bus lane
  - Give confidence level & justification for the decision
- **Success Criteria:**
  - How do we measure "being effective" ?
- **Analytics Goal:**
  - Discover what factors most influence success?
  - Given attributes of a new bus lane, predict if it will be successful

Possible Success Criteria

- Increase in passengers on buses using the bus lane?
- Increase in bus punctuality (less clumping)?
- Shortened bus journey times?
- Reduced traffic along the route?
- All of the above? (e.g. use a weighted success function)

How much increase or decrease constitutes success?

Example: Identifying Data Requirements

- How effective are existing bus lanes?
- What distinguishes effective from ineffective lanes?

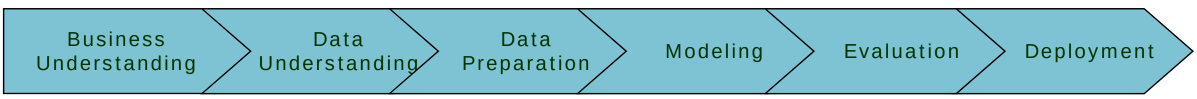
For each Bus Lane…	Data required to answer (suggestions)
Is there an increase in bus riders?	• Number of riders boarding at each bus stop • List of bus stops in each bus route
Is there shortened bus journey times?	
Is there an increase in bus punctuality? (less clumping?)	
Is there reduced traffic along the route?	
Does effectiveness depend on day, time?	
Was effectiveness sustained?	

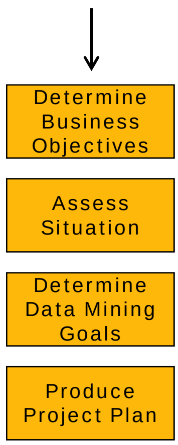
Is this enough?

Example: Identifying Data Requirements

For Each Bus Lane ….	Data required to answer (suggestions)
Is there an increase in bus riders?	• Number of riders boarding at each bus stop at each time of day/DOW and data for before & after the bus lane is introduced • List of bus stops in each bus route.
Is there shortened bus journey times?	• Time taken by bus to get from one bus stop to the next per route (before & after)
Is there an increase in bus punctuality? (less clumping?)	• Bus arrival times at each bus stop (scheduled and actual; before and after) • Time between bus arrivals at each stop per route (before & after)
Is there reduced traffic along the route?	• Congestion figures along route (before & after)
Does effectiveness depend on day, time?	• All of above, but broken down by TOD, DOW
Was effectiveness sustained?	• Above data for many months

Putting it all together: Generating a Plan





- List the stages to be executed in the project
  - Include their duration, resources required, inputs, outputs
  - Analyze dependencies between the stages and the risks
  - Include the ***initial selection*** of tools and techniques - No need to specify the exact algorithms/methods at this time
  - Where possible, make explicit the large-scale ***iterations*** in the data mining process—for example, repetitions of the modeling and evaluation phases
- Be ***Agile***…
  - Try to work in short iterations. Each iteration generates a prototype working system which is tested and then refined and improved over time – using test results and feedback from stakeholders

A Less Agile Internship Project Plan

Understand → **Prepare** → **Model** → **Test** → **Deploy**

Understand	Prepare	Model	Test	Deploy
Aug ->Sep	Oct->Nov	Dec->Jan	Feb	Mar

A More Agile Internship Project Plan

Increasing Understanding ->

Increasing Data Preparation ->

Improving Models ->

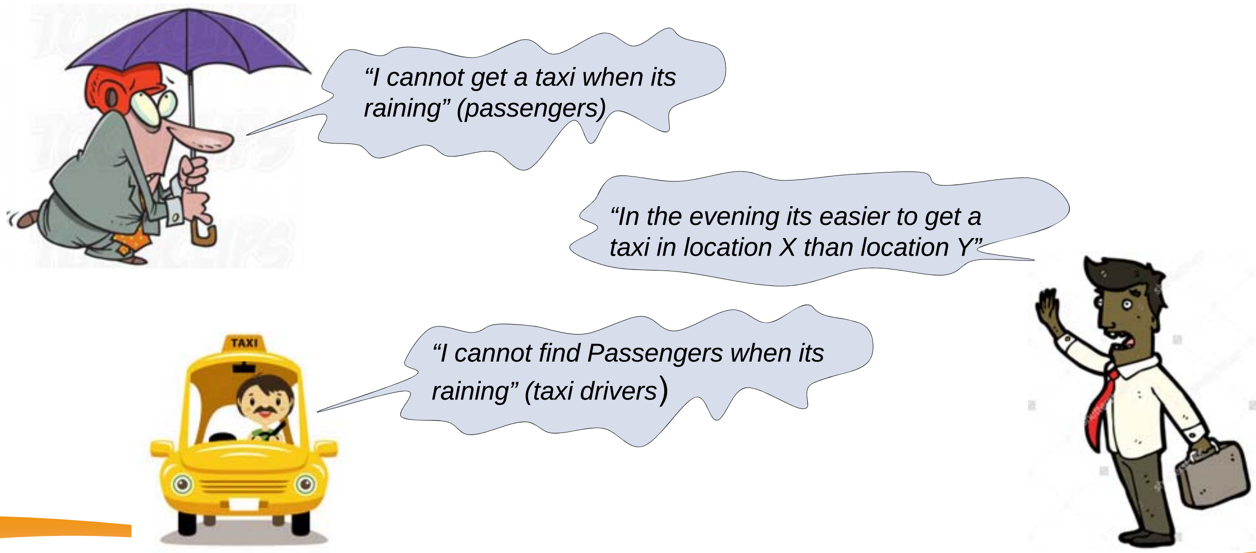
Lab Testing ->

Production Testing ->

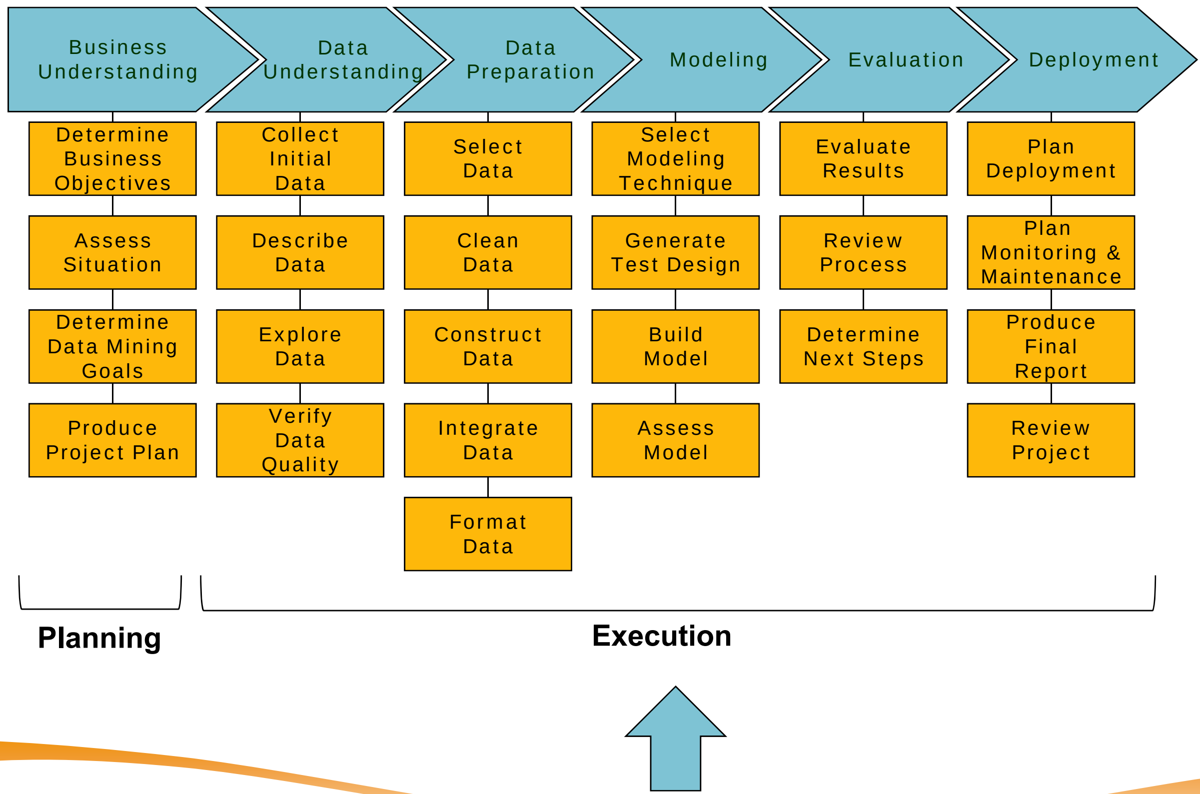
Iteration 1   Iteration 2 . . . . Iteration N

Identifying Data Requirements - Workshop

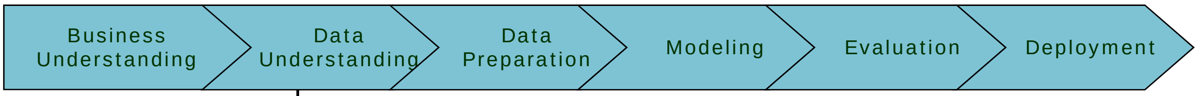
- How might we use data analytics to improve taxi availability?
- Investigating feedback & known issues can be a starting point
- What data is required to validate these claims below? How could it be derived?
- How would you use the data to validate the claim? (what analytics approach?)

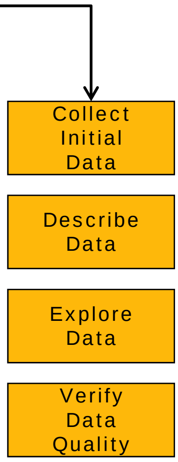


Project Execution



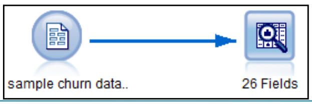
Data Understanding / Exploration

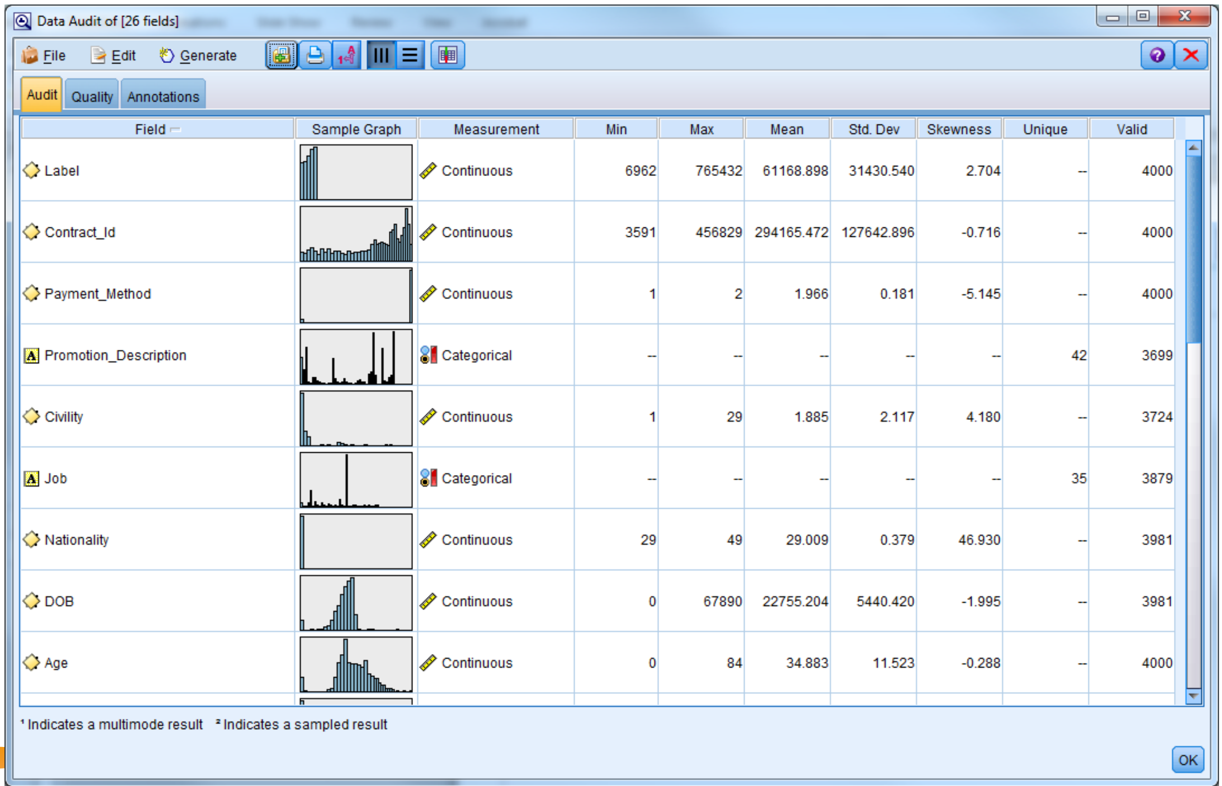




- A useful starting point is the Data Audit. Audit goals include:
  - Is the data adequate?
  - Is it what you expect?
  - Does it look sensible?
  - What are the data quality issues? (What cleaning is required?)
- Data Exploration is more concerned with analysis and discovery (can also be done on the prepared data)
  - Find answers to questions asked
  - Make recommendations
  - Find Insights
  - Data visualization is a key tool

SPSS Modeler Data Audit Node





Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Label		Continuous	6962	765432	61168.898	31430.540	2.704	--	4000
Contract_Id		Continuous	3591	456829	294165.472	127642.896	-0.716	--	4000
Payment_Method		Continuous	1	2	1.966	0.181	-5.145	--	4000
Promotion_Description		Categorical	--	--	--	--	--	42	3699
Civility		Continuous	1	29	1.885	2.117	4.180	--	3724
Job		Categorical	--	--	--	--	--	35	3879
Nationality		Continuous	29	49	29.009	0.379	46.930	--	3981
DOB		Continuous	0	67890	22755.204	5440.420	-1.995	--	3981
Age		Continuous	0	84	34.883	11.523	-0.288	--	4000

[1] Indicates a multimode result [2] Indicates a sampled result

SPSS Modeler Data Audit Node

Data Audit of [26 fields]

File Edit Generate

Audit | Quality | Annotations

Complete fields (%): 42.31% Complete records (%): 0%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Label	Continuous	0	1	None	Never	Fixed	100	4000	0	0	0	0
Contract_Id	Continuous	0	0	None	Never	Fixed	100	4000	0	0	0	0
Payment_Method	Continuous	0	136	None	Never	Fixed	100	4000	0	0	0	0
Promotion_Description	Categorical	--	--	--	Never	Fixed	92.475	3699	0	301	301	0
Civility	Continuous	189	5	None	Never	Fixed	93.1	3724	276	0	0	0
Job	Categorical	--	--	--	Never	Fixed	96.975	3879	0	121	121	0
Nationality	Continuous	0	3	None	Never	Fixed	99.525	3981	19	0	0	0
DOB	Continuous	121	2	None	Never	Fixed	99.525	3981	19	0	0	0
Age	Continuous	133	0	None	Never	Fixed	100	4000	0	0	0	0
Age_Band	Continuous	0	18	None	Never	Fixed	97.05	3882	118	0	0	0
Gender	Continuous	0	18	None	Never	Fixed	96	3840	160	0	0	0
Credit_Score	Categorical	--	--	--	Never	Fixed	61.225	2449	0	1551	1551	0
Tariff_Plan	Continuous	0	0	None	Never	Fixed	99.95	3998	2	0	0	0
Num_Active_VAS	Continuous	0	11	None	Never	Fixed	99.8	3992	8	0	0	0
Num_Inactive_VAS	Continuous	2	7	None	Never	Fixed	100	4000	0	0	0	0
Service_DataFax	Continuous	0	11	None	Never	Fixed	96.65	3866	134	0	0	0
Service_Voicemail	Continuous	1	9	None	Never	Fixed	55	2200	1800	0	0	0
Service_SMS	Categorical	--	--	--	Never	Fixed	0.45	18	0	3982	3982	0
Cust_Activation_Date	Continuous	0	15	None	Never	Fixed	99.825	3993	7	0	0	0
Cust_Contact	Continuous	83	17	None	Never	Fixed	99.525	3981	19	0	0	0
Cust_Contact_Compl...	Continuous	0	19	None	Never	Fixed	100	4000	0	0	0	0
Num_active_VAS_pre...	Continuous	0	0	None	Never	Fixed	100	4000	0	0	0	0
Num_inactive_VAS_pr...	Continuous	0	0	None	Never	Fixed	100	4000	0	0	0	0
Cust_Contact_prev_...	Continuous	5	13	None	Never	Fixed	100	4000	0	0	0	0
Cust_Contact_Compl...	Continuous	1	10	None	Never	Fixed	100	4000	0	0	0	0
Churn_Flag	Continuous	0	0	None	Never	Fixed	100	4000	0	0	0	0

OK

Data Audit in R

Requires library(Hmisc)

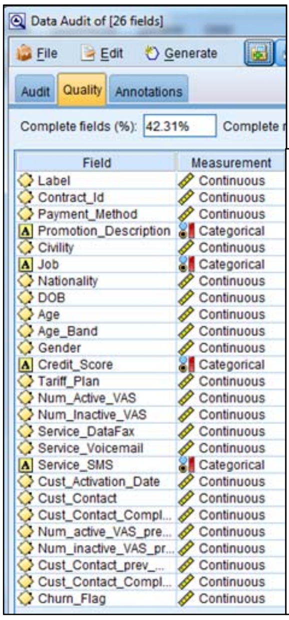
```
> data <- read.csv("HP sample churn data.csv")
> contents(data)

Data frame:data 4000 observations and 26 variables    Maximum # NAs:4000


                                  Levels Storage  NAs
Label                                    integer    0
Contract_Id                              integer    0
Payment_Method                           integer    0
Promotion_Description                 42 integer    0
Civility                                 integer  276
Job                                   35 integer    0
Nationality                              integer    0
DOB                                      integer    0
Age                                      integer    0
Age_Band                                 integer  119
Gender                                   integer  162
Credit_Score                             integer 1558
Tariff_Plan                              integer    0
Num_Active_VAS                           integer    0
Num_Inactive_VAS                         integer    0
Service_DataFax                          integer  135
Service_Voicemail                        integer 1807
Service_SMS                              logical 4000
Cust_Activation_Date                     integer    0
Cust_Contact                             integer    0
Cust_Contact_Complaints                  integer    0
Num_active_VAS_prev_month                integer    0
Num_inactive_VAS_prev_month              integer    0
Cust_Contact_prev_month                  integer    0
Cust_Contact_Complaints_prev_month       integer    0
Churn_Flag                               integer    0

+---------------------+--------------------------------------------------------------------------------+
|Variable             |Levels                                                                          |
+---------------------+--------------------------------------------------------------------------------+
|Promotion_Description|E - baya Clubs,E - HLBank,E - HLBank v2,E - HLBank v3,E - KL Mutual,E - MBf MasterCard |
|                     |E - Standard Chartered,E - Standard Chartered v2,X - MBGSP1 -b Gp st H/P1,X - MCSDP2 -Cosway St'kst PL2 |
|                     |X - MPBARO -b Agent /Rem's,X - MPBSAD -b S'ger/Agt,X - MPCCPL -Corp Conversion,X - MPCIT2 -Mut cb PL 2 |
|                     |X - MPCITI -Mut cb PL,X - MPCMP1,X - MPCMP2,X - MPCMP3 -Cosway Members PL3,X - MPCOR1 -Corp Company PL 1|
|                     |X - MPCOR2 -Corp Company PL 2,X - MPCSD1 -Cosway Stikst  PL1,X - MPCSDP,X - MPMAP3 -Major Account PL3 |
|                     |X - MPMBSP,X - MPMIPO,X - MPMJPL,X - MPMSHP -SR1 st Purch,X - MPMSIP -Mut st IPO PL,X - MPOMNI |
|                     |X - MPPBP2 -Public Plan 2,X - MPPBP3 -Public Plan 3,X - MPPBP5 -Public Plan 5,X - MPPBPL |
|                     |X - MPTV3P -TV3 H/P Plan,X - MTKPPL,Z - 1999 Ultimate Mobile Pack,Z - Double Bonus,Z - Festive Promo |
|                     |Z - Free test - 2000,Z - None,Z - S2D '99                                       |
+---------------------+--------------------------------------------------------------------------------+
|Job                  |Advertising / Media,Agriculture, Forestry, Fishing,Banks & Financial Institutions |
|                     |Business & Technical Services,Computer & Communications,Construction / Housing,Consultants |
|                     |Consulting & Security Company,Education,Engineering,Engineering Architecture,Government / Agencies |
|                     |Housewife,Import/Export,Insurance Services,Legal Services,Manufacturing,Medical & Health |
|                     |Mining & Quarrying,Others,Professional,Real Estate,Restaurant/Hotel,Retail Trade,Security |
|                     |Sole Proprietor,Student,Telecommunication,Tour & Hotel,Transport/Store,Transportation,Travel & Tour |
|                     |Wholesale Trade,Wholesaler & Retailer                                           |
+---------------------+--------------------------------------------------------------------------------+
```

```
> summary(data)
     Label          Contract_Id       Payment_Method                  Promotion_Description     Civility     
 Min.   :  6962   Min.   :  3591   Min.   :1.000   Z - S2D '99                  : 659      Min.   : 1.000  
 1st Qu.: 35136   1st Qu.:209189   1st Qu.:2.000   X - MPPBP5 -Public Plan 5    : 641      1st Qu.: 1.000  
 Median : 62644   Median :337566   Median :2.000   E - HLBank v2                : 459      Median : 1.000  
 Mean   : 61169   Mean   :294166   Mean   :1.966   Z - 1999 Ultimate Mobile Pack: 421      Mean   : 1.885  
 3rd Qu.: 84511   3rd Qu.:404133   3rd Qu.:2.000   X - MPCIT2 -Mut cb PL 2      : 306      3rd Qu.: 2.000  
 Max.   :765432   Max.   :456829   Max.   :2.000                                : 301      Max.   :29.000  
                                                  (Other)                       :1213      NA's   :276     
                              Job        Nationality        DOB             Age           Age_Band         Gender     
 Others                        :1893   Min.   :29.00   Min.   :    0   Min.   : 0.00   Min.   :1.000   Min.   :1.000  
 Business & Technical Services : 587   1st Qu.:29.00   1st Qu.:20767   1st Qu.:28.00   1st Qu.:3.000   1st Qu.:1.000  
 Manufacturing                 : 244   Median :29.00   Median :23850   Median :34.00   Median :4.000   Median :1.000  
 Construction / Housing        : 163   Mean   :29.01   Mean   :22749   Mean   :34.92   Mean   :4.444   Mean   :1.332  
 Banks & Financial Institutions: 123   3rd Qu.:29.00   3rd Qu.:26212   3rd Qu.:42.00   3rd Qu.:6.000   3rd Qu.:2.000  
 Education                     : 122   Max.   :49.00   Max.   :67890   Max.   :84.00   Max.   :8.000   Max.   :4.000  
 (Other)                       : 868                                                   NA's   :119     NA's   :162    
  Credit_Score    Tariff_Plan     Num_Active_VAS  Num_Inactive_VAS Service_DataFax Service_Voicemail Service_SMS   
 Min.   :170.0   Min.   : 4.00   Min.   : 0.00   Min.   : 0.000   Min.   :1.000   Min.   :1.000     Mode:logical  
 1st Qu.:229.0   1st Qu.: 4.00   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.:1.000   1st Qu.:1.000     NA's:4000     
 Median :250.0   Median :12.00   Median :17.00   Median : 2.000   Median :1.000   Median :1.000                   
 Mean   :248.8   Mean   :41.84   Mean   :11.09   Mean   : 8.538   Mean   :1.239   Mean   :1.249                   
 3rd Qu.:267.0   3rd Qu.:85.00   3rd Qu.:20.00   3rd Qu.:20.000   3rd Qu.:1.000   3rd Qu.:1.000                   
 Max.   :349.0   Max.   :85.00   Max.   :29.00   Max.   :29.000   Max.   :3.000   Max.   :3.000                   
 NA's   :1558                                                     NA's   :135     NA's   :1807                    
 Cust_Activation_Date  Cust_Contact    Cust_Contact_Complaints Num_active_VAS_prev_month Num_inactive_VAS_prev_month
 Min.   : 1234        Min.   : 0.00   Min.   :0.00000         Min.   : 0.00             Min.   : 0.000             
 1st Qu.:35742        1st Qu.: 0.00   1st Qu.:0.00000         1st Qu.: 0.00             1st Qu.: 0.000             
 Median :36214        Median : 0.00   Median :0.00000         Median :17.00             Median : 2.000             
 Mean   :36043        Mean   : 0.36   Mean   :0.01375         Mean   :10.78             Mean   : 8.829             
 3rd Qu.:36441        3rd Qu.: 0.00   3rd Qu.:0.00000         3rd Qu.:20.00             3rd Qu.:20.000             
 Max.   :98765        Max.   :13.00   Max.   :6.00000         Max.   :29.00             Max.   :29.000             

 Cust_Contact_prev_month Cust_Contact_Complaints_prev_month   Churn_Flag 
 Min.   :  0.0000        Min.   :0.0000                     Min.   :0.0  
 1st Qu.:  0.0000        1st Qu.:0.0000                     1st Qu.:0.0  
 Median :  0.0000        Median :0.0000                     Median :0.5  
 Mean   :  0.3105        Mean   :0.0095                     Mean   :0.5  
 3rd Qu.:  0.0000        3rd Qu.:0.0000                     3rd Qu.:1.0  
 Max.   :115.0000        Max.   :2.0000                     Max.   :1.0  
```



```
> describe(data)
data 

 26  Variables      4000  Observations
---------------------------------------------------------------------------------------------------------------
Label 
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95 
    4000        0     4000        1    61169    34057    13538    19417    35136    62644    84511    99796   100444 

Value       10000  20000  30000  40000  50000  60000  70000  80000  90000 100000 770000
Frequency     250    370    375    336    336    457    433    462      8    972      1
Proportion  0.062  0.092  0.094  0.084  0.084  0.114  0.108  0.116  0.002  0.243  0.000
---------------------------------------------------------------------------------------------------------------
Contract_Id 
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95 
    4000        0     4000        1   294165   142150    44617    81219   209189   337566   404133   432088   439970 

lowest :   3591   3702   3985   4454   4461, highest: 456452 456692 456720 456787 456829
---------------------------------------------------------------------------------------------------------------
Payment_Method 
       n  missing distinct     Info     Mean      Gmd 
    4000        0        2    0.099    1.966   0.0657 

Value          1     2
Frequency    136  3864
Proportion 0.034 0.966
---------------------------------------------------------------------------------------------------------------
Promotion_Description 
       n  missing distinct 
    4000        0       42 

lowest :                         E - baya Clubs       E - HLBank           E - HLBank v2        E - HLBank v3       
highest: Z - Double Bonus        Z - Festive Promo    Z - Free test - 2000 Z - None             Z - S2D '99         
---------------------------------------------------------------------------------------------------------------
Civility 
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95 
    3724      276       14    0.706    1.885    1.448        1        1        1        1        2        3        9 

Value          1     2     3     5     6     7     9    10    11    12    15    21    22    29
Frequency   2455   678   372     1    13    11    91    83    13     2     1     1     1     2
Proportion 0.659 0.182 0.100 0.000 0.003 0.003 0.024 0.022 0.003 0.001 0.000 0.000 0.000 0.001
---------------------------------------------------------------------------------------------------------------
```

Are the assigned Data Types correct?



Field	Measurement
Label	Continuous
Contract_Id	Continuous
Payment_Method	Continuous
Promotion_Description	Categorical
Civility	Continuous
Job	Categorical
Nationality	Continuous
DOB	Continuous
Age	Continuous
Age_Band	Continuous
Gender	Continuous
Credit_Score	Categorical
Tariff_Plan	Continuous
Num_Active_VAS	Continuous
Num_Inactive_VAS	Continuous
Service_DataFax	Continuous
Service_Voicemail	Continuous
Service_SMS	Categorical
Cust_Activation_Date	Continuous
Cust_Contact	Continuous
Cust_Contact_Compl...	Continuous
Num_active_VAS_pre...	Continuous
Num_inactive_VAS_pr...	Continuous
Cust_Contact_prev_...	Continuous
Cust_Contact_Compl...	Continuous
Churn_Flag	Continuous

SPSS modeler

```
                               Levels Storage
Label                                 integer
Contract_Id                           integer
Payment_Method                        integer
Promotion_Description              42 integer
Civility                              integer
Job                                35 integer
Nationality                           integer
DOB                                   integer
Age                                   integer
Age_Band                              integer
Gender                                integer
Credit_Score                          integer
Tariff_Plan                           integer
Num_Active_VAS                        integer
Num_Inactive_VAS                      integer
Service_DataFax                       integer
Service_Voicemail                     integer
Service_SMS                           logical
Cust_Activation_Date                  integer
Cust_Contact                          integer
Cust_Contact_Complaints               integer
Num_active_VAS_prev_month             integer
Num_inactive_VAS_prev_month           integer
Cust_Contact_prev_month               integer
Cust_Contact_Complaints_prev          integer
Churn_Flag                            integer
```

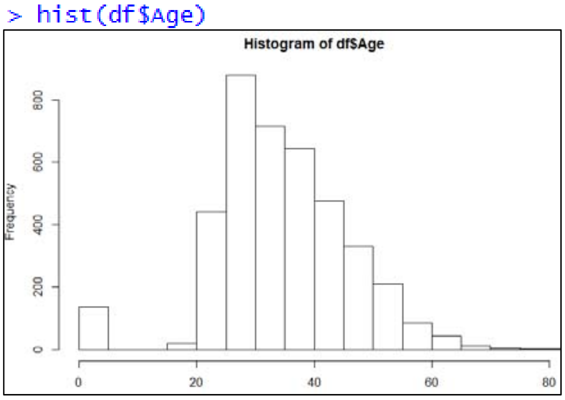
R

```
> head(df)
  Label Contract_Id Payment_Method     Promotion_Description Civility
1 12345      211353              2 X - MPPBP5 -Public Plan 5        1
2  6962      250588              2   X - MPCIT2 -Mut cb PL 2        1
3  7036       71782              2 X - MPPBP3 -Public Plan 3        1
4  7082       12717              2 X - MPPBP2 -Public Plan 2        1
5  7093       57437              1 X - MPPBP2 -Public Plan 2        1
6  7141      264101              2   X - MPCIT2 -Mut cb PL 2        1
                         Job Nationality   DOB Age Age_Band Gender
1                     Others          29 67890  40        6      1
2                     Others          29 24257  34        4      1
3                     Others          29 19778  46        6      1
4       Wholesaler & Retailer          29 25206  31        4     NA
5 Business & Technical Services          29 26243  28        2      1
6                     Others          29 21927  40        3      1
  Credit_Score Tariff_Plan Num_Active_VAS Num_Inactive_VAS Service_DataFax
1           NA          53             17                6               1
2           NA           4             18                0               1
3           NA           4             22                0               1
4           NA          85              0               27               1
5           NA          53             17                6               1
6           NA           4             21                0               1
  Service_Voicemail Service_SMS Cust_Activation_Date Cust_Contact
1                NA          NA                 9876            0
2                 1          NA                35846            0
3                NA          NA                35203            0
4                NA          NA                35150            1
5                 3          NA                35076            0
6                 1          NA                35891            0
  Cust_Contact_Complaints Num_active_VAS_prev_month
1                       0                        17
2                       0                        18
3                       0                        22
4                       0                         0
5                       0                        17
6                       0                        21
  Num_inactive_VAS_prev_month Cust_Contact_prev_month
1                           6                       0
2                           0                       0
3                           0                       0
4                          27                       2
5                           6                       0
6                           0                       0
  Cust_Contact_Complaints_prev_month Churn_Flag
1                                  0          1
2                                  0          1
3                                  0          1
4                                  0          0
5                                  0          1
6                                  0          0
>
```

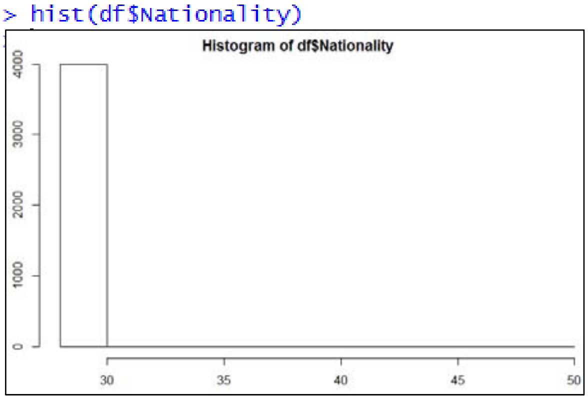
Categorical variable are often stored as numbers

Investigating Data Types

```
>
> hist(df$Age)
```

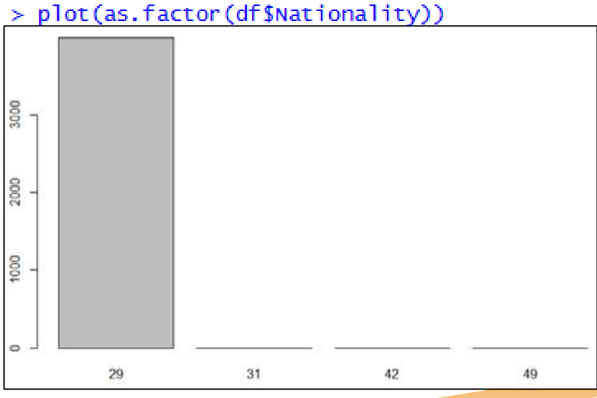



```
>
> hist(df$Nationality)
```




```
> table(df$Nationality)

  29   31   42   49 
3997    1    1    1 
>
```

```
>
> plot(as.factor(df$Nationality))
```



Major Tasks in Data Preparation*

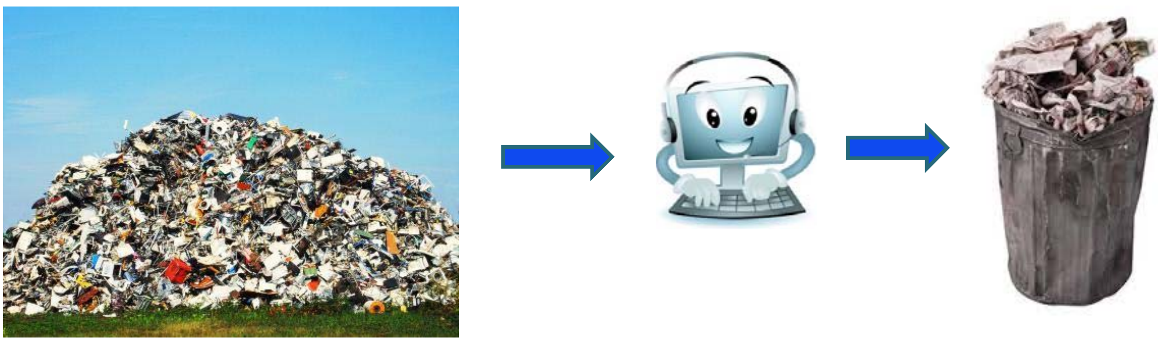
Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

Select Data
Clean Data
Construct Data
Integrate Data
Format Data

1. Data cleaning
   – Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
2. Data integration
   – Integration of multiple databases, handle duplications, inconsistencies etc.
3. Data transformation / construction
   – Enhance the data through normalization, aggregation, feature creation etc.
4. Data reduction / selection
   – Select data relevant to the business problem
   – Reduce the volume of the data (as required) while ensuring the same or similar analytical results

*The order of the steps can be varied

Step1: Data Cleaning



G.I.G.O.

Data Cleaning

- Data may not be perfectly collected, or collected with the right purpose.
- Many reasons exist for data to be dirty:
  - Data entry errors
  - Misplaced decimal points
  - Inherent error in counting or measuring devices
  - External factors, etc.
- Data exploration can discover anomalous patterns, leading to the questioning of data quality
  - E.g. categories with very low frequency counts ➔ mistyping?
  - Name and addresses recorded in multiple ways in data integrated from multiple sources (can be up to 20~30 variations)
  - Missing data

Data Cleaning Tasks

- Data cleaning tasks
  - Handle missing values
  - Handle noisy / erroneous data
  - Handle outliers
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Values

- Common feature of any dataset
- Various reasons:
  - Information not available
  - Lost data / accidentally deleted
  - Purposefully left out with a reason
- Missing does not always imply an empty/blank value. There may be a value entered in the data that signifies missing
  - E.g. *"9999", "1 Jan 1900", "*", "?", "#", "$"*, etc
- The presence of missing values in data can make problems for the modeling tools.

Handling Missing Values

- Ignore attributes that have majority of values missing?
- Ignore data records with missing values?
  - Throwing away data ~ but this is bad if you do not have much data!
  - Especially poor when the percentage of missing values per attribute varies considerably – one attribute (which may not even be important) with few values could cause the whole data to be discarded!

Gender	Children	Salary	Bought PEP
M	-	29,000	Y
M	-	65,000	Y
F	2	26,500	Y
M	-	47,000	Y
F	-	15,000	N
-	1	23,000	N
F	-	36,000	N

What should we do here?

Handling Missing Values

- Data Imputation - fill in the missing values automatically
  – Guiding Principle: Avoid adding bias and distortion to the data
  – Understand why the data is missing can help guide the imputation
  – Often a missing value means zero or the default value. E.g. for '*rainfall*' variable, a missing value may mean no rain on that day ➔ 0

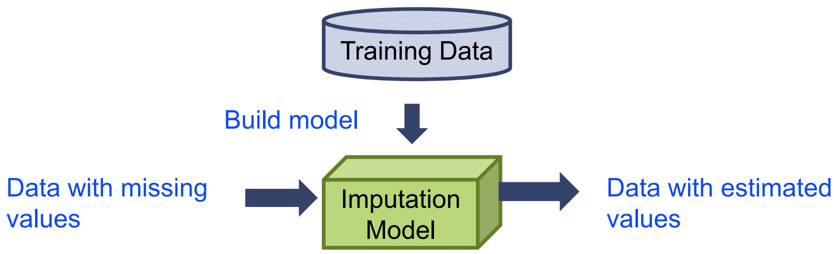
- Common Options
  - A global **constant** : e.g., "unknown" or 0 (zero)
    Easy, but modeling algorithms may mistakingly treat "unknown" as a concept
  - The **attribute mean** (or median, mode)
    Simple and quick though not always satisfactory
  - The **attribute mean** for all samples belonging to the **same class**
    Often a better estimate than attribute mean

Gender	Children	Salary	Bought PEP
M	1	29,000	Y
M	0	65,000	Y
F	2	-	Y
M	0	47,000	Y
F	-	15,000	N
-	1	23,000	N
F	1	36,000	N

What should we do here?

Data Imputation

- Train a prediction model (e.g. regression model, decision tree) to predict the most probable value
  - Use variables containing values to estimate the variable with missing values
  - Can produce good estimates.
  - Need training data and additional modeling

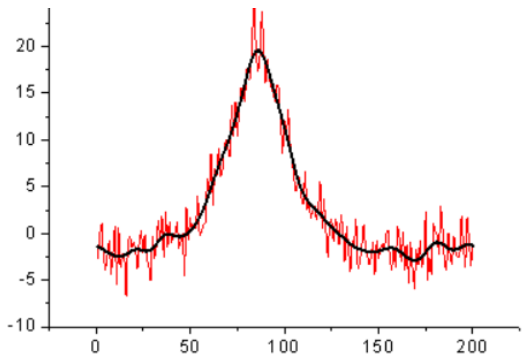


Noisy / Erroneous Data

- **Noise:** random error or variance in a measured variable
- Incorrect attribute values may have been entered due to
  - Measurement error: faulty (or inaccurate) data collection instruments
  - Data entry problems
  - Data transmission problems
  - Inconsistency in naming convention
  - Others....

Noise handling Methods

- Binning
  - Sort and bin data, use bin means, medians etc
- Curve/Line Fitting
  - Fitting the data into regression functions
- Ensemble methods
  - Averaging the results from multiple models



Outliers

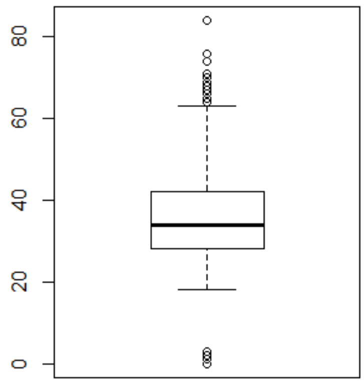
- Observations that "*deviate so much from other observations as to arouse suspicion that it was generated by a different mechanism*". (Hawkins, 1980)
- Appearing at the maximum or minimum end of a variable, skewing or distorting the distribution
  - E.g. extreme weather conditions on a particular day, a very wealthy person financially very different from the rest of the population, etc.

Handling Outliers

- Outliers may be errors ***or*** they may be valid data!
  - Can be rare, unusual, infrequent events we are interested in.
  - They should be identified for further investigation.
  - E.g. frauds in income tax, insurance, banking, etc.
- Otherwise, outliers usually should be removed to avoid adversely affecting the modeling result (though some algorithms, like random forests and support vector machines can be robust to outliers)

Identifying Outliers

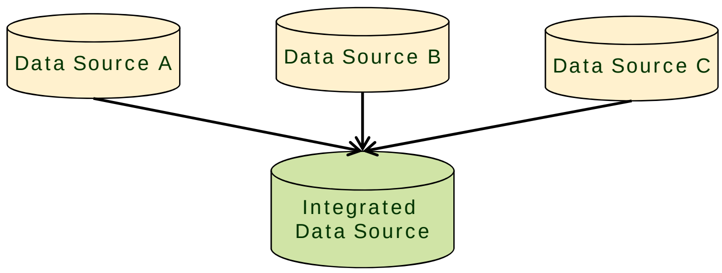
- Statistical tests for variance
- Clustering
- Human inspection
- Others…




```
> boxplot(df$Age)
```


Step2: Data Integration

- Combining data from different sources into a coherent store



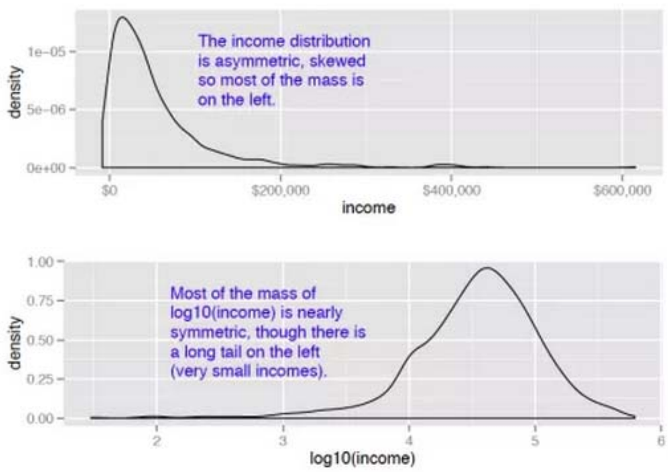
- Duplication & Redundancy
  - Same attribute may have different names in different databases (e.g. tenure, length of service)
  - One attribute may be derived from another in a different database (e.g. monthly and annual revenue)
  - Same user may be identified differently in different databases (e.g. "John Smith" vs "Smith, J.")
- Inconsistency & Data Value Conflicts
  - Same attribute may occur in different databases but with different values for the same entity
    e.g. Ben's account age in database1 is 24 months , while in database2 it's 2 years
  - Possible reasons: different representations, different scales, different time zones
    e.g., Metric vs. British units

Step3: Data Transformation

- **Smoothing:** remove noise from data
- **Log Transformation:** remove skew
- **Square Root Transformation:** remove skew
- **Normalization:** scaled to fall within a small, specified range
- **Aggregation:** summarization , data reduction
- **Generalization:** concept hierarchy climbing
- **Category to number conversion:** handling categorical variables
- **Feature construction:** data enhancement
- Others….

Log Transformations

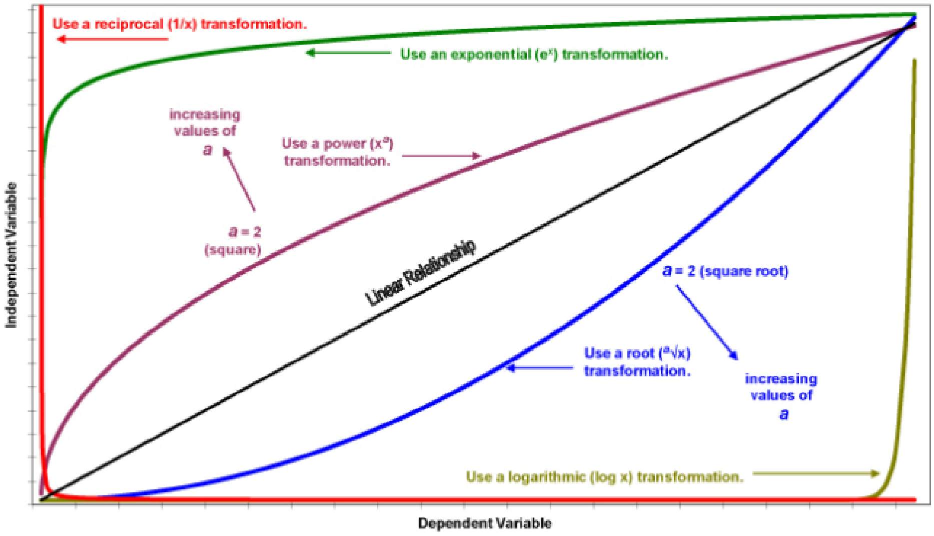
- Log Transformation
  - Makes a skewed attribute more symmetric
  - Reduces the magnitudes
  - Common bases 10, 2, *e* *(which base to use is often not important)*



- Incomes, customer value, account or purchase sizes—are commonly encountered sources of skewed distributions in data science applications.
- Often they are log-normally distributed: the log of the data is normally distributed

Log Transformations

If a data relationship looks like one of these curves,
try using a transformation of the independent variable
to make the relationship linear.



- https://statswithcats.wordpress.com/2010/11/21/fifty-ways-to-fix-your-data/

Data Normalization

- Reduces outlier distortion and enhances linear predictability
- Ensure all variables have approximately the same scale
  - E.g. variable *Age* vs *Income*: a distance of 10 "years" may be more significant than a distance of \$1000, yet \$1000 swamps 10 when they are added in calculating distance
- Normally re-center and rescale the data to be around zero, in the range from 0 to 1, etc.
- Common Methods.....

$$v' = \frac{v - min_A}{max_A - min_A}$$

Min-max scaling

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Z-score scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max($|v'|$)<1

Decimal scaling

Handling Categorical Data

- Many modeling methods require numerical inputs
  – One major exception is decision tree methods
- How to convert categories into numbers without introducing an unintended ordering?
- E.g. Which of these is the best mapping?
  - Small ->1
  - Medium -> 2
  - Large -> 3

  - Small ->3
  - Medium -> 2
  - Large -> 1

  - Small ->2
  - Medium -> 3
  - Large -> 1
- What about this?
  - Yishun->1
  - Clementi -> 2
  - Tuas-> 3
  - Queensway -> 4

Handling Categorical Data

- How to handle…
  – Marital status = single, married, divorced, widowed?
- Could convert to…
  – Marital status = 0,1,2,3 where
  0 = single, 1=married, 2=divorced, 3=widowed
- Better to create four new T/F variables
  – Single = 0,1
  – Married = 0,1
  – Divorced = 0,1
  – Widowed = 0,1

- Caution:
  – For visualisation and decision tree models, it's best to leave as one field called "marital status" with values = single, married, divorced, widowed

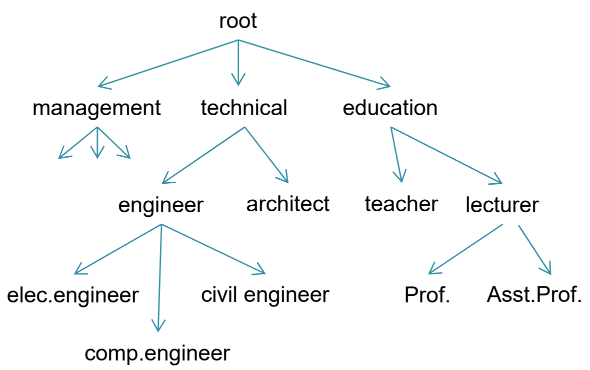
Handling Categorical Data

- **If** there is no obvious ordering within the categories then converting to a series of binary (1 => true and 0 => false) inputs is preferable
- This is often also called "one-hot" encoding or "dummy" variable encoding
- Example

Obs.	Colour	Colour_Red	Colour_Green	Colour_Blue
1	Green	0	1	0
2	Blue	0	0	1
3	Blue	0	0	1
4	Red	1	0	0
5	Green	0	1	0
6	Red	1	0	0

Handling Categorical Data

- Simplify categorical variables that have too many categories before doing binarisation
- Simple grouping may help
  - E.g. transform states into groups: western, eastern etc.
- If a concept hierarchy exists then categories can be merged by climbing the hierarchy
- E.g.…….



Gender	Profession	Bought PEP
M	teacher	Y
M	professor	Y
F	Asst. professor	Y
M	Civil engineer	N
F	Comp.engineer	N
F	Elec. engineer	N
M	architect	N

Gender	Profession	Bought PEP
M	education	Y
M	education	Y
F	education	Y
M	technical	N
F	technical	N
F	technical	N
M	technical	N

Feature Construction

- Decomposing compound features into simpler components, e.g....

ID	Product Holdings	Purchased Service
1.	ProdA + ProdC	Y
2.	ProdB + ProdC	N
3.	ProdA + ProdD	N
4.	ProdB + ProdD	Y
...		

ProdA	ProdB	ProdC	ProdD	Svc
1. 1	0	1	0	Y
2. 0	1	1	0	N
3. 1	0	0	1	N
4. 0	1	0	1	Y
...				

Feature Construction

- Deriving a value that is more useful / making something more explicit
- E.g.

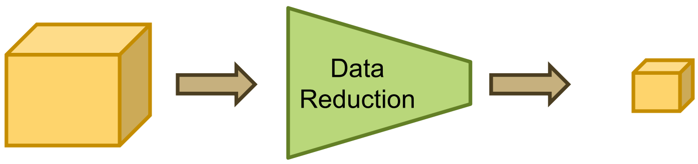
ID	Cost per unit	Units purchased
1.	10	10
2.	15	5
3.	8	8
4.	10	5

ID	Cost per unit	Units purchased	Total $ Revenue
1.	10	10	100
2.	15	5	75
3.	8	8	64
4.	10	5	50

- Other examples
  - Age = current date - date of birth
  - Area = length * width

Step4: Data Reduction

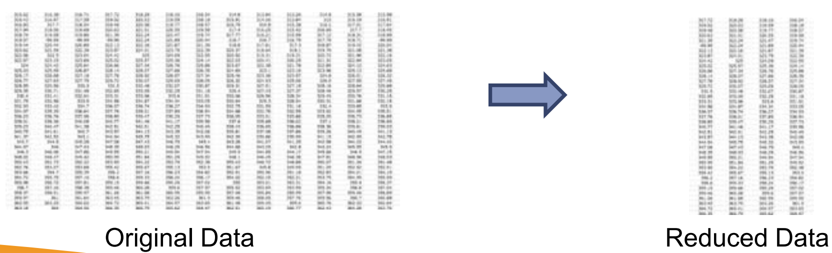
- Complex data analytics may take a very long time to run on the complete data set
- Data Reduction
  - Obtain a reduced representation of the data set that is much smaller in volume yet produces the same (or almost the same) analytical results
- Data Reduction Strategies
  - Dimensionality reduction—reduce the number of attributes
  - Numerosity reduction – reduce by finding alternate, smaller data representations
    - Parametric methods: - fit data into models, store model parameters, discard the data
    - Non-parametric methods - histograms, clustering, sampling



Dimensionality Reduction

- **Feature Selection** (attribute subset selection)
  - Selecting the most relevant attributes

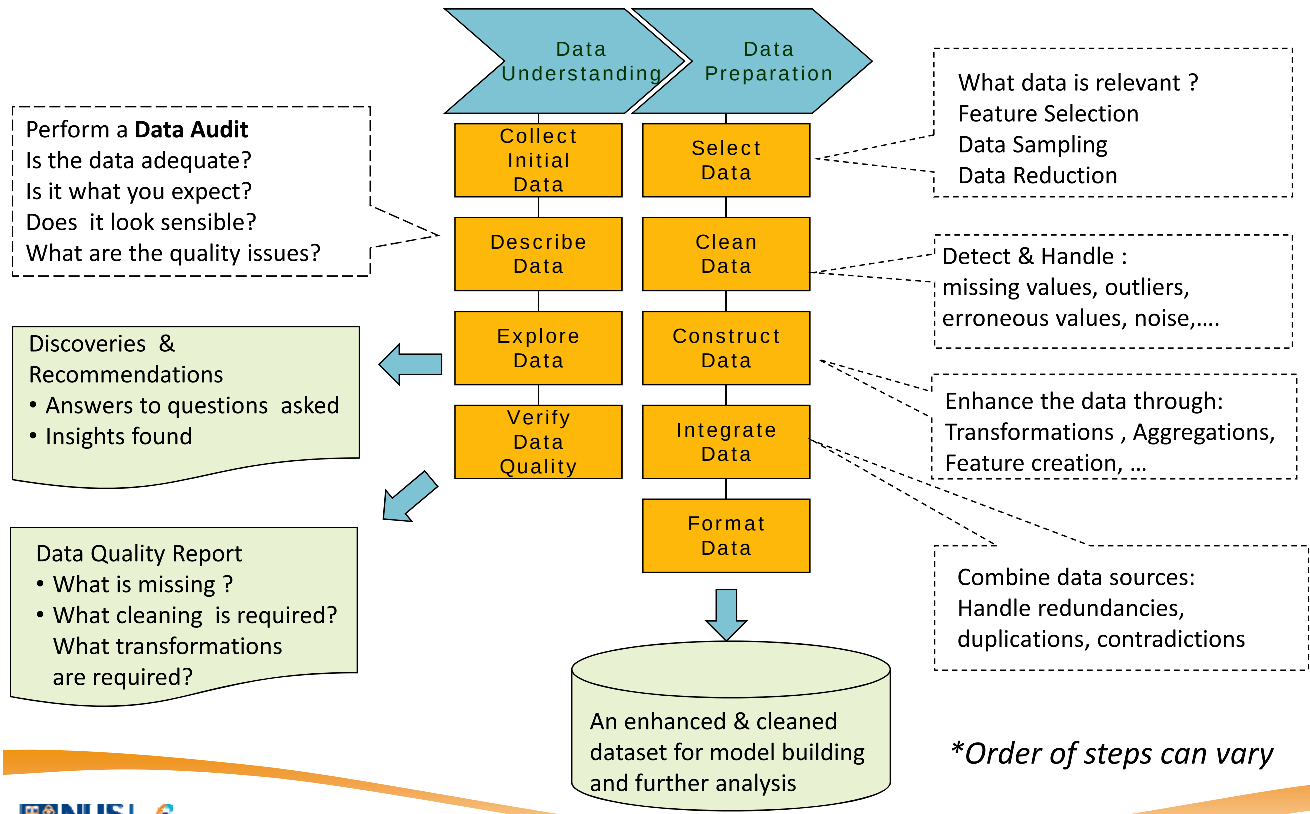
- **Feature Extraction**
  - Combining attributes into a new reduced set of features



Feature Extraction

- Also attribute reduction process by combining the original attributes
- Leading to a much smaller and richer set of attributes
- Methods exist which work well for linear between-variable relationships
  - Principle component analysis
  - Factor analysis

Data Preparation Summary



**Order of steps can vary*