

Vancouver SAS User's Group

# Data Cleansing Process

## Common Practices

Peter Hruby, Sr. Manager  
Credit Risk Management



## Data Cleansing Process - Common Practices

---

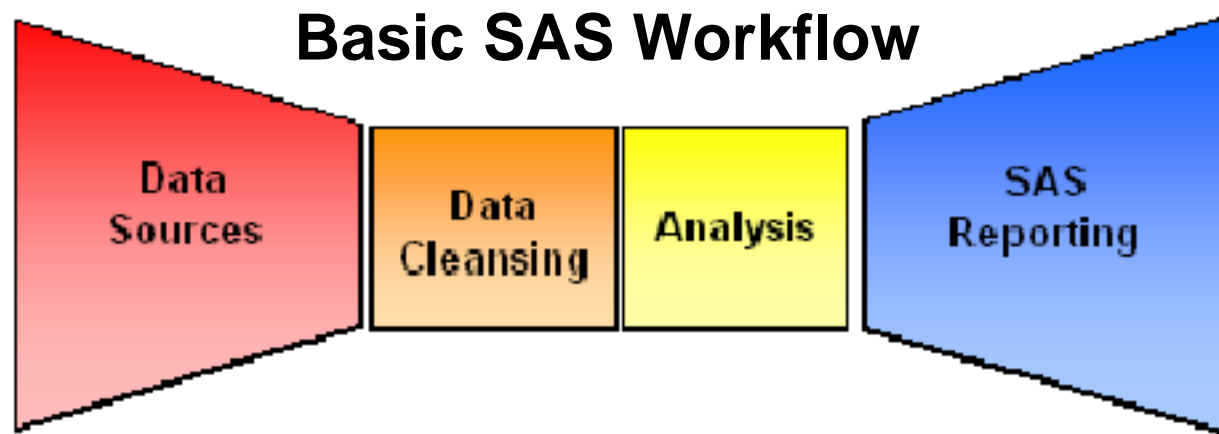
**We're going to talk about**

- ▶ **Analytic process**
- ▶ **Factors causing data integrity issues**
- ▶ **Character variables**
- ▶ **Numeric variables**
- ▶ **Working with data errors and outliers**
- ▶ **Missing values**
- ▶ **Duplicates – analyst nightmare!**
- ▶ **Faster methods for data cleansing**
- ▶ **Common analytic errors**
- ▶ **Data cleansing Top 10 recommendations**
- ▶ **Documentation and training courses**

## Data Cleansing Process - Common Practices

---

### Analytic Process



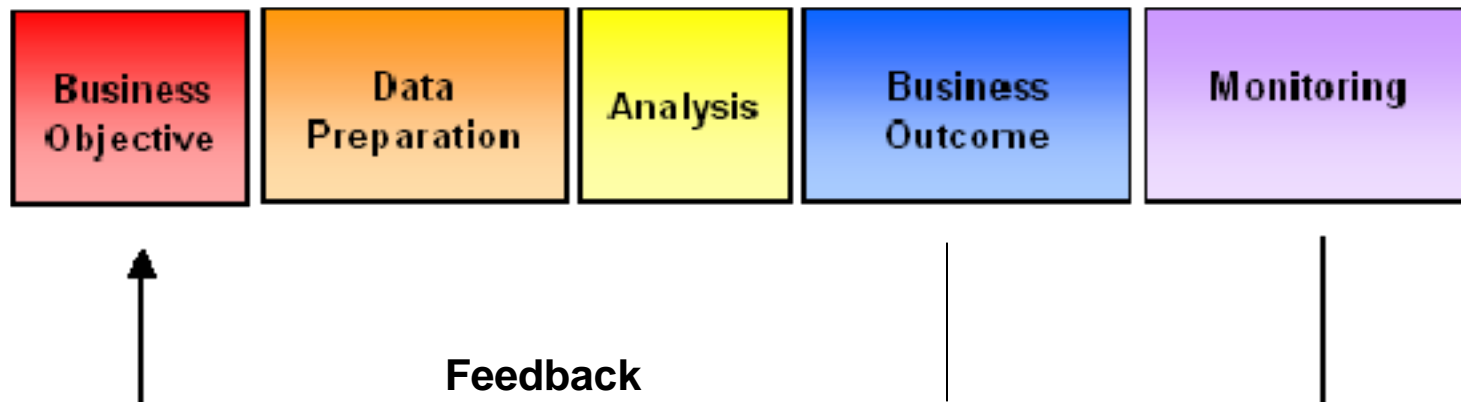
Data preparation and data cleansing is one of the most important steps in any analytic process. From the simplest analysis to the most complex model, the quality of the data is key to the success of the project.

## Data Cleansing Process - Common Practices

---

### Analytic Process

#### Business Flow



- ▶ Analyst devotes up to 85% of total time to data cleaning and preparation
- ▶ Data cleansing involves looking for and handling data errors, outliers, missing values, while controlling duplicity issues during SAS execution.

## Data Cleansing Process - Common Practices

---

### Analytic Process

#### Business Objective

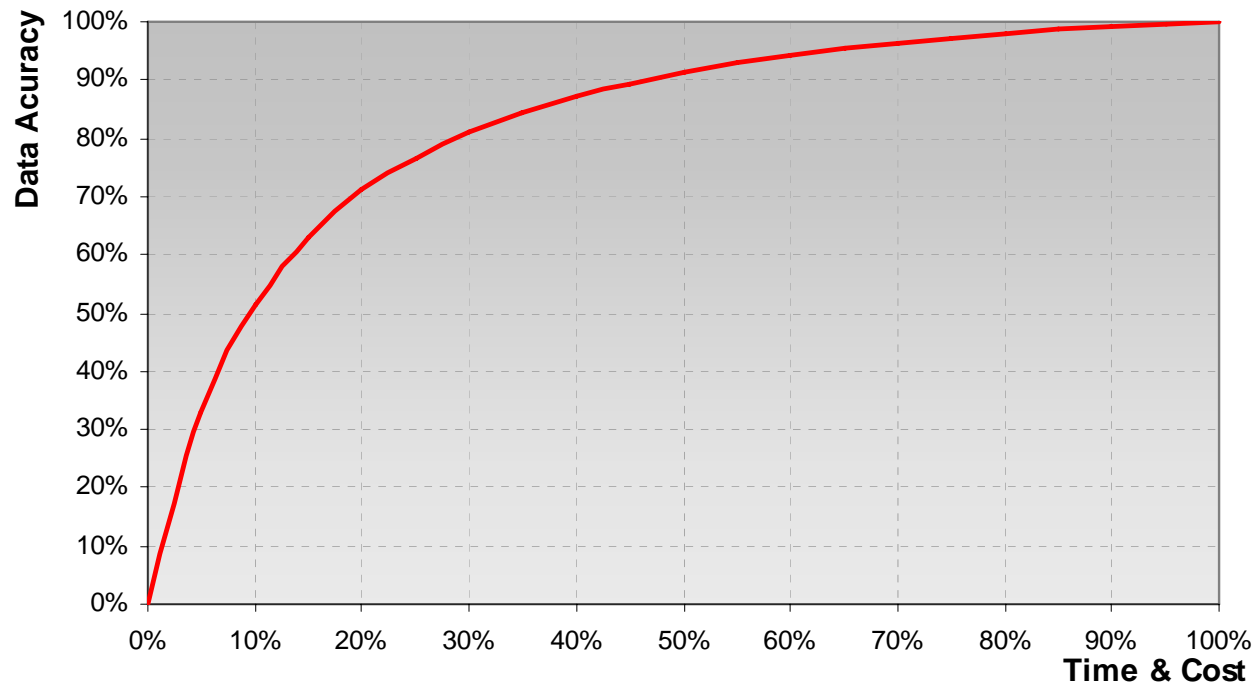
- ▶ Ad-hoc vs. Regular reporting?
- ▶ Simple statistics vs. higher statistics?
- ▶ Small or large population?
- ▶ Any time constraints?
- ▶ Hardware Resources?

# Data Cleansing Process - Common Practices

---

Analytic Process

## Data Cleansing Trade-off curve



## Data Cleansing Process - Common Practices

---

### Factors causing data integrity issues

- ▶ Business Environment is constantly evolving (business grows...data grows)
- ▶ Input errors! (30-40% of the data is inputted manually)
- ▶ New data variables (new products and services generates new data)
- ▶ Personnel changes (career development, rotation of human resources)
- ▶ New technology implementation (new database, new platform)

## Data Cleansing Process - Common Practices

---

### Character variables

- ▶ Represents mainly categorical data  
Analyst's task is either correct data entries or assign special category for further analysis
- ▶ Possibility to use external data sources to verify housed data (postal codes, address verification, phone numbers area codes)
- ▶ Reduce categorical data into common groups (Employer, Job description, Goods).
- ▶ PROC PRINT
- ▶ PROC FREQ
- ▶ PROC FORMAT

## Data Cleansing Process - Common Practices

---

### Numeric variables

- ▶ Qualitative Data  
Used for segmentation or classification, ie. Gender, Province
- ▶ PROC PRINT, PROC FREQ
- ▶ PROC FORMAT, PROC TABULATE
  
- ▶ Quantitative Data  
Used as independent variables in various statistics
- ▶ PROC PRINT, PROC FREQ
- ▶ PROC FORMAT, PROC TABULATE
- ▶ PROC MEANS, PROC UNIVARIATE,
- ▶ PROC RANK

## Data Cleansing Process - Common Practices

---

### Numeric variables

- ▶ Nominal Data  
Represents categories with no relative importance  
(Gender, Marital status)
- ▶ Ordinal Data  
Represents categories with relative importance  
(Delinquency, Overlimit)
- ▶ Continuous Data  
Most common data with relative importance  
(Salaries, Purchases, Scores)

## Data Cleansing Process - Common Practices

---

### Working with data errors and outliers

- ▶ Outlier is a single or low frequency occurrence of the value of the variable that is far from the mean as well as the majority of the other values within variable
- ▶ Fixing an outlier is one of the most time consuming part of the cleaning process
- ▶ There is no exact process to identify an outlier
- ▶ Common sense and good logic, as well as knowing your data will lead to some solution.
- ▶ For monetary variables, capping rule is a good method to fix outliers

## Data Cleansing Process - Common Practices

---

### Working with data errors and outliers

- ▶ PROC UNIVARIATE is the most efficient procedure for identifying outliers
- ▶ When used with TRIM option, Proc Univariate will generate mean and standard deviation, excluding proportion maximum and minimum values from analysis.
- ▶ For normal distributions:  
Mean +/- 1.96\*St.Dev contains approximately 95% of all variables  
Mean +/- 2.57\*St.Dev contains approximately 99% of all variables
- ▶ Categorical data should have error values either deleted or replaced with correct value

## Data Cleansing Process - Common Practices

---

### Missing values

- ▶ Missing values are present in majority of SAS data set
- ▶ It is vitally important to carefully inspect the SAS Log for any indication of missing values
- ▶ PROC FREQ, PROC PRINT

How to encounter a missing value:

- ▶ Missing Raw data
- ▶ Invalid raw value causes a missing SAS value
- ▶ Computation of new variable can cause a missing SAS value (division by 0,  $\ln(0)$ , arithmetic operation using missing value)

# Data Cleansing Process - Common Practices

## Missing values

```
data p325pri_&yymm.;
  infile " p325pri_&yymm..txt" TRUNCOVER lrecl=838 firstobs=2;
  input @140 LST_CONT_ACT_DATE yymmdd6.
        @632 MEMO_DATE yymmdd6.;
run;
```

NOTE: Invalid data for LST\_CONT\_ACT\_DATE in line 2 140-145.  
NOTE: Invalid data for MEMO\_DATE in line 2 632-637.

```
RULE:      -+-----1-+-----2-+-----3-+-----4-+-----5-+-----6-+-----7-+-----
-8-----+
      2  XXXXXXXXXXXXXXXXXXXX          101115001032755700          99901      13      0NNNN
1561      0
      87      0      0      0      0      0      31125      31030      31112          0          0      0
0NN
      173          .000000          .000000          .000000          0      0
0
      259          .000000          000000      031000S20QSUP          8          1942.990000
      345      1196.010000      0          .000000          .000000NS33      31126      0
      431          .000000 NNN      31125MW01      QSUP      S40NYNNNNNN00          0          0
      517          .000000          .000000          .000000          0          31125
      603          .000000          .000000          .000000          .000000          .000000
      689          .000000          .000000          .000000          .000000          .000000
      775          .000000          000N      NN          PY      1      31030
LST_CONT_ACT_DATE=. MEMO_DATE=. _ERROR_=1 _N_=1
```

# Data Cleansing Process - Common Practices

---

## Missing values

```
data p325pri;
infile " p325pri_&yymm..txt" TRUNCOVER lrecl=838 firstobs=2;
input @140 tmp_LST_CONT_ACT_DATE $6.
      @632 tmp_MEMO_DATE $6.;

if tmp_LST_CONT_ACT_DATE = '      0' then LST_CONT_ACT_DATE = .;
else LST_CONT_ACT_DATE = input(tmp_LST_CONT_ACT_DATE, yymmdd6.);

if tmp_MEMO_DATE = '      0' then MEMO_DATE = .;
else MEMO_DATE = input(tmp_MEMO_DATE, yymmdd6.);
run;
```

## Data Cleansing Process - Common Practices

---

### Missing values

- ▶ Create new category or substitute the value?
- ▶ Categorical data should have missing values defined as separate category only

Following techniques are commonly used for substitution:

- ▶ Simple substitution (Mode, Median, Mean)  
PROC UNIVARIATE
- ▶ Class substitution using correlation analysis or decision trees  
PROC TABULATE
- ▶ Linear regression  
PROC REG

## Data Cleansing Process - Common Practices

---

### Duplicates – An analyst nightmare!

- ▶ Check for duplicities after every merge or update  
PROC SORT, PROC SQL, PROC FREQ
- ▶ Avoid cartesian merges (Exception is one record with many).
- ▶ Begin with master dataset which has no duplicates
- ▶ Use one-side joins to add and new variables. (Depends on business....)
- ▶ Merge vs. Update statement. Is there a difference?

```
%macro sort(dsn,var);  
  proc sort data    = &dsn  
            out     = &dsn._sort  
            dupout  = dupes nodupkey;  
    by &var.;  
  run;  
%mend;
```

## Data Cleansing Process - Common Practices

---

### Faster methods for data cleansing

- ▶ Use macros, it's easy and decreases processing time

```
%macro print(source,fobs,obs,cond);  
  proc print data = &source. (firstobs = &fobs. obs = &obs.);  
    &cond;  
  run;  
%mend;
```

```
%macro freq(source,var,cond);  
  proc freq data = &source.;  
    table &var. / nocol norow nopercnt;  
    &cond;  
  run;  
%mend;
```

```
%macro unvar(source,var,cond);  
  proc univariate data = &source. plot;  
    var &var.;  
    &cond;  
  run;  
%mend;
```

## Data Cleansing Process - Common Practices

---

### Faster methods for data cleansing

- ▶ Insert macro code into autoexec.sas, macros will be available all the time you open SAS session
- ▶ Use custom formats for continuous variables

```
proc format;  
  value score  
    0 - .1 = '(90%-100%]'  
    .1 <- .2 = '(80%- 90%]'  
    .2 <- .3 = '(70%- 80%]'  
    .3 <- .4 = '(60%- 70%]'  
    .4 <- .5 = '(50%- 60%]'  
    .5 <- .6 = '(40%- 50%]'  
    .6 <- .7 = '(30%- 40%]'  
    .7 <- .8 = '(20%- 30%]'  
    .8 <- .9 = '(10%- 20%]'  
    .9 <- 1 = '( 0%- 10%]';  
run;
```

## Data Cleansing Process - Common Practices

---

### Faster methods for data cleansing

- ▶ When dealing with large datasets, use STOP statement for faster processing

```
data dnbpifp (keep = acct_num BBDLPC BBGHCL);
merge DNBPIFP.dw_&_31x_01b_yyyymm. (in      = a
                                     keep    = acct_num BBDLPC BBGHCL
                                     rename  = (BBDLPC = BBDLPC_24))
      DNBPIFP.dw_&_31x_07b_yyyymm. (keep    = acct_num BBDLPC
                                     rename  = (BBDLPC = BBDLPC_30));

length BBDLPC $30.;
by acct_num;
if a;
length = length(trimn(BBDLPC_30));
if length > 20 then BBDLPC1_30 = substr(BBDLPC_30,19,length-18);
BBDLPC = BBDLPC_24||trimn(BBDLPC1_30);
if _n_ = 15000 then stop;
run;
```

## Data Cleansing Process - Common Practices

---

### Common errors

- ▶ Not paying close attention to black and blue SAS notes.  
(SAS went to a new line while reading input data; Variable has been converted from .. to..; There is a merge of the variables with different length)
- ▶ Ignoring green warning messages
- ▶ Merging numeric data with character data
- ▶ Merging data with different length format
- ▶ Omission to check new variable results
- ▶ Duplicity
- ▶ Avoid dropping unnecessary variables from data sets
- ▶ Checking for data errors

## Data Cleansing Process - Common Practices

---

### Data cleansing Top 10

- ▶ Data cleansing is a repeating process
- ▶ Thorough understanding of the business objective.
- ▶ Knowing your data
- ▶ Paying attention to details
- ▶ Common sense (does the results makes any sense?)
- ▶ Using cleansing methods you're comfortable with!
- ▶ Implement cleansing at the source
- ▶ Share information with your colleagues
- ▶ Use one data source across organization
- ▶ Check for duplicities during data step

## Data Cleansing Process - Common Practices

---

### Recommended Documentation

#### Book

Ron Cody

### Cody's Data Cleaning Techniques using SAS

Web link: <http://www.sas.com/apps/pubscat/bookdetails.jsp?pc=61703>

#### Training Course

### Data Cleaning Techniques

Web link: <http://support.sas.com/training/us/crs/bdct.html>