What’s really in your data?
A quick and easy utility to analyze any SAS dataset

SESUG Author, Michele Laughlin, Catalyst Technologies, Snellville, GA
SESUG Co-Author, Thom Clay, Catalyst Technologies, Snellville, GA
SESUG Co-Author, Jamila Ervin, Catalyst Technologies, Snellville, GA

ABSTRACT
Many companies produce Quality Assurance reports daily in order to provide the highest quality of service and keep information up to date. The reports may consist of general demographic information, statistical data, cost analysis, date validation, etc. In order to produce all of these reports, programmers may generate code to create each of these analyses. The effort in creating separate utilities can be time consuming, expensive and an inefficient use of company resources. When Catalyst Technologies’ management realized the number of hours expanded on creating code for these Quality Assurance reports, they came up with an answer. This paper discusses that solution, an easy to use utility that combines a variety of data to create numerous reports usable by programmers and non-programmers alike.

TECHNICAL OVERVIEW
The utility uses a PROC CONTENTS statement to read in all of the variables on the input dataset. It will then evaluate each individual variable in the dataset to determine which report to produce based on the variable type. If a numeric variable is discovered, then a PROC UNIVARIATE will be performed. If a character variable is found, then a PROC FREQUENCY will be performed. The utility also looks for a date variable and, if found, will perform a PROC FREQUENCY using the year of the date. There are additional reports that can be produced from this utility. Examples include age, date of birth frequency and reporting of values that fall outside of a specified range.

%macro FreqUniv(indsn);
%inc "G:\Macros\DateIt.sas";
%inc "G:\Macros\DateRpt.sas";
%inc "G:\Macros\FreqIt.sas";
%inc "G:\Macros\HighVol.sas";
%inc "G:\Macros\UnivIt.sas";

Proc contents data=&indsn
    out=contents (keep=name type length format formatl formatd
    label nobs) nopr;
run;

filename sasparm "E:\Customer_Regions\Sample_Customer\ sastest.txt";

data contents;
set contents(rename=(type=vtype));
file sasparm;
if type  = 1 and format  in ('DATE', 'DDMMYY', 'JULIAN', 'MMDDYY', 'MMYY',
'MONYY', 'WORDDATE', 'YYMMD', 'YYMON')
    or
    informat in ('DATE', 'DDMMYY', 'JULIAN', 'MMDDYY', 'MMYY',
'MONYY', 'WORDDATE', 'YYMMD', 'YYMON')
then do;
dtflag = 'Y';
After the PROC CONTENTS is performed, the utility decides what to do with the data it has been given. If a numeric variable is found, the PROC UNIVARIATE will be run by calling the macro `univit`.

```sas
*** Macro to run univariates based on &var ***;
%macro univit(var);
title2 "Univariate Statistics of &var from &indsn";
proc univariate data=&indsn;
  var &var;
  output out=temp_unv
  NOBS = NOBS
  N = N
  MEAN = MEAN
  SUM = SUM
  STD = STD
  MAX = MAX
  MIN = MIN
  MEDIAN = MEDIAN
  VAR = VAR
  MODE = MODE
  P10 = P10
  P90 = P90
  nmiss = nmiss;
run;
%if &freqout ~= %then %do;
data temp_unv;
  length custid $4 datatype $7 varname $20 value $50 count percent cumpct rundate 8;
  set temp_unv;
  varname  = "&var";
  rundate  = datetime();
  value    = 'Number of Missing values';
  count    = nmiss;
  percent  = round(nmiss/nobs,.0001);
  output;
  value    = 'Number of Nonmissing values';
  count    = n;
  percent  = round(n/nobs,.0001);
  output;
  value    = 'The Sum';
  count    = sum;
  percent  = .;
  output;
  value    = 'The Largest Value';
  count    = max;
run;
%mend univit;
```

%univit(custds.sample);
This produces an output report containing useful statistics such as standard deviations, means, moments, sum weights, base statistical measures, quantiles, maximum, mean and minimums.

**OUTPUT REPORT**

Sample Customer
Univariate Statistics of DAYSSPLY from custds.sample

The UNIVARIATE Procedure
Variable: DAYSSPLY (Days Supply)

Moments
N          32000        Sum Weights  32000
Mean       27.7229063 Sum Observations 887133
Std Deviation 23.1790809 Variance  537.26979
Skewness   2.43620053 Kurtosis  22.0769411
Uncorrected SS  41786001 Corrected SS  17192096
Coeff Variation 83.6098519 Std Error Mean  0.129575

Basic Statistical Measures

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>27.72291</td>
</tr>
<tr>
<td>Median</td>
<td>30.00000</td>
</tr>
<tr>
<td>Mode</td>
<td>30.00000</td>
</tr>
</tbody>
</table>

Interquartile Range 20.00000

Quantiles (Definition 5)

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>520</td>
</tr>
<tr>
<td>99%</td>
<td>90</td>
</tr>
<tr>
<td>95%</td>
<td>90</td>
</tr>
<tr>
<td>90%</td>
<td>34</td>
</tr>
<tr>
<td>75% Q3</td>
<td>30</td>
</tr>
<tr>
<td>50% Median</td>
<td>30</td>
</tr>
<tr>
<td>25% Q1</td>
<td>10</td>
</tr>
<tr>
<td>10%</td>
<td>4</td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
</tr>
<tr>
<td>1%</td>
<td>0</td>
</tr>
<tr>
<td>0% Min</td>
<td>-90</td>
</tr>
</tbody>
</table>

Extreme Observations

<table>
<thead>
<tr>
<th>Value</th>
<th>Obs</th>
<th>Value</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>-90</td>
<td>19358</td>
<td>336</td>
<td>7406</td>
</tr>
<tr>
<td>-90</td>
<td>15038</td>
<td>365</td>
<td>11454</td>
</tr>
<tr>
<td>-90</td>
<td>9103</td>
<td>365</td>
<td>11455</td>
</tr>
<tr>
<td>-90</td>
<td>9102</td>
<td>520</td>
<td>13029</td>
</tr>
<tr>
<td>-90</td>
<td>2451</td>
<td>520</td>
<td>13031</td>
</tr>
</tbody>
</table>

If a character variable is found, the utility will determine the length of the field. If the length is less than three then a PROC FREQUENCY will be performed on all of the observations by calling the macro 'freqit'.

%macro freqit(var);
  title2 'Frequency distribution of &var from &indsn';
  proc freq data=&indsn;
    table &var / missing out=temp_frq;
  run;
  %if &freqout ~= %then %do;
    data temp_frq;
    length custid $4 datatype $7 varname $20 value $50 count percent cumpct rundate 8;
    set temp_frq;
    varname = "&var";
    value   = &var;
    percent = round(percent/100,.0001);
cumcnt + count;
cumpct = round(cumcnt/\subset1,.0001);

rundate = datetime();
drop &var cumcnt;
retain cumcnt 0;
format rundate datetime. count comma12. percent cumpct percent9.2;
label varname = 'Variable Name'
  value = 'Value'
  count = 'Count'
  percent = 'Percent'
  cumpct = 'Cumulative Percent'
  rundate = 'Run Date'
;
run;

proc append base=freqout.&freqpfx._freqtbl data=temp_frq;
run;
%end;
%mend freqit;

If the length is greater than three, then the macro will limit the number of observations output by performing a PROC FREQUENCY on the first 100 observations in the dataset. If the utility finds that there are more than three observations, the macro 'highvol' is invoked.

*** Macro to run high volumn frequency for &var ***;
%macro highvol(var);
    proc freq data=&indsn noprint order=freq;
        table &var / missing out=temp_hv;
    data temp_hv;
        set temp_hv(obs=100);
    proc sort data=temp_hv;
        by &var;
    data temp_hv;
        set temp_hv;
        percent = round(percent/100,.0001);
        cumcnt + count;
        cumpct = cumcnt;
    run;
    title2 "Top 100 from frequency distribution of &var from &indsn";
    proc print data=temp_hv split='*' noobs;
        var &var count cumcnt percent cumpct;
        format percent cumpct percent9.2;
        label count = 'Frequency'
            cumcnt = 'Cumulative*Frequency'
            percent = 'Percent*of Total'
            cumpct = 'Cumulative*Percent';
    run;
%end highvol;

The output report below shows a frequency distribution of the sample variable. The number of observations output can be changed to whatever you wish. The default used in this example is 2.

**OUTPUT REPORT**

Sample Customer
Top 10 from frequency distribution of SAMPVAR
The report below shows the FREQ procedure for the variable DOS summarized by year.

**OUTPUT REPORT**

Sample Customer
Frequency distribution of DOS by year from custds.Sample

The FREQ Procedure
Date of Service
### CONCLUSION

The FREQUNIV utility is a general purpose data analysis tool for use with any SAS data set. The program dynamically determines the report set based on individual variable types and reduces valuable resource time by eliminating redundant programming for multiple data sources.

<table>
<thead>
<tr>
<th>DOS</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>27</td>
<td>0.08</td>
<td>27</td>
<td>0.08</td>
</tr>
<tr>
<td>2006</td>
<td>31973</td>
<td>99.92</td>
<td>32000</td>
<td>100.00</td>
</tr>
<tr>
<td>Name</td>
<td>Address</td>
<td>Address</td>
<td>Address</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td></td>
</tr>
<tr>
<td>Thom Clay</td>
<td>2386 Clower Street</td>
<td>2386 Clower Street</td>
<td>2386 Clower Street</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Suite C201</td>
<td>Suite C201</td>
<td>Suite C201</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Snellville, GA 30078</td>
<td>Snellville, GA 30078</td>
<td>Snellville, GA 30078</td>
<td></td>
</tr>
<tr>
<td></td>
<td>770-982-8022 ext. 164</td>
<td>770-982-8022 ext. 126</td>
<td>770-982-8022 ext. 133</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="mailto:tclay@catalysttech.com">tclay@catalysttech.com</a></td>
<td><a href="mailto:jervin@catalysttech.com">jervin@catalysttech.com</a></td>
<td><a href="mailto:mlaughlin@catalysttech.com">mlaughlin@catalysttech.com</a></td>
<td></td>
</tr>
</tbody>
</table>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., in the USA and other countries. © Indicates USA registration.
**FREQUNIV FLOWCHART**

**Inputs:**
- `%indsn.` = Input Dataset
- `%freqobs` = Max number of obs to report
- `%freqout` = Output path of reports dataset
- `%macro vars` = Limit vars to process
- `%macro rgvars` = Vars to perform Range process

**FREQUNIV Process**

```
PROC CONTENTS
```

**Determine Variable Type**

```
Determine Variable Type
```

**PROC UNIV** (%macro univit)

```
If Numeric
```

**PROC FREQ** (%macro freqit)

```
If Character
```

**PROC FREQ** (%macro highvol)

```
If Date
```

**Determine Variable Length**

```
If Length <= 3
```

**PROC FREQ** (%macro freg)

```
If Length > 3
```

**PROC FREQ** Top 100 (%macro highvol)

**Range Report** (%macro rrange)

**PROC CONTENTS DSET**

```
Limit Vars? (%macro vars)
```

**Range Vars? (%macro rgvars)**

```
If DOB
```

**PROC FREQ** w/YEAR Format (%macro dateit)

```
If Character
```

**Age Report** (%macro daterept)

```
If Numeric
```

**Report**

**Inputs:**
- `%indsn.` = Input Dataset
- `%freqobs` = Max number of obs to report
- `%freqout` = Output path of reports dataset
- `%macro vars` = Limit vars to process
- `%macro rgvars` = Vars to perform Range process