

Unknown Knowns: Database Construction from Unknown Files and Variables

William Klein

Knowns and unknowns

File names

Known

Unknown

Variable names

Known

Unknown

Assumptions

Windows

External text files

Names of data files are not known

Number of data files is not known

Names of the variables are not known

Each data file has the same number of variables

Objectives

Retail chain sales report

Assemble a SAS file to concatenate (join together) all the required external text files

Assemble a SAS file to contain one record (row) for each observation

Known file names and variables

```
data known_file;
```

```
  infile 'C:\knowns.dat';
```

```
  input hour jobs @@ @;
```

```
run;
```

```
proc print;
```

```
run;
```

The log confirms your choice of file name and location

NOTE: The infile 'C:\knowns.dat' is:

Filename=C:\DATA\knowns.dat,

RECFM=V,LRECL=256,File Size (bytes)=125,

Last Modified=October 07, 2010 14:02:57

Create Time=October 07, 2010 14:02:57

NOTE: 2 records were read from the infile
'C:\knowns.dat'.

Searching for unknown files and folders

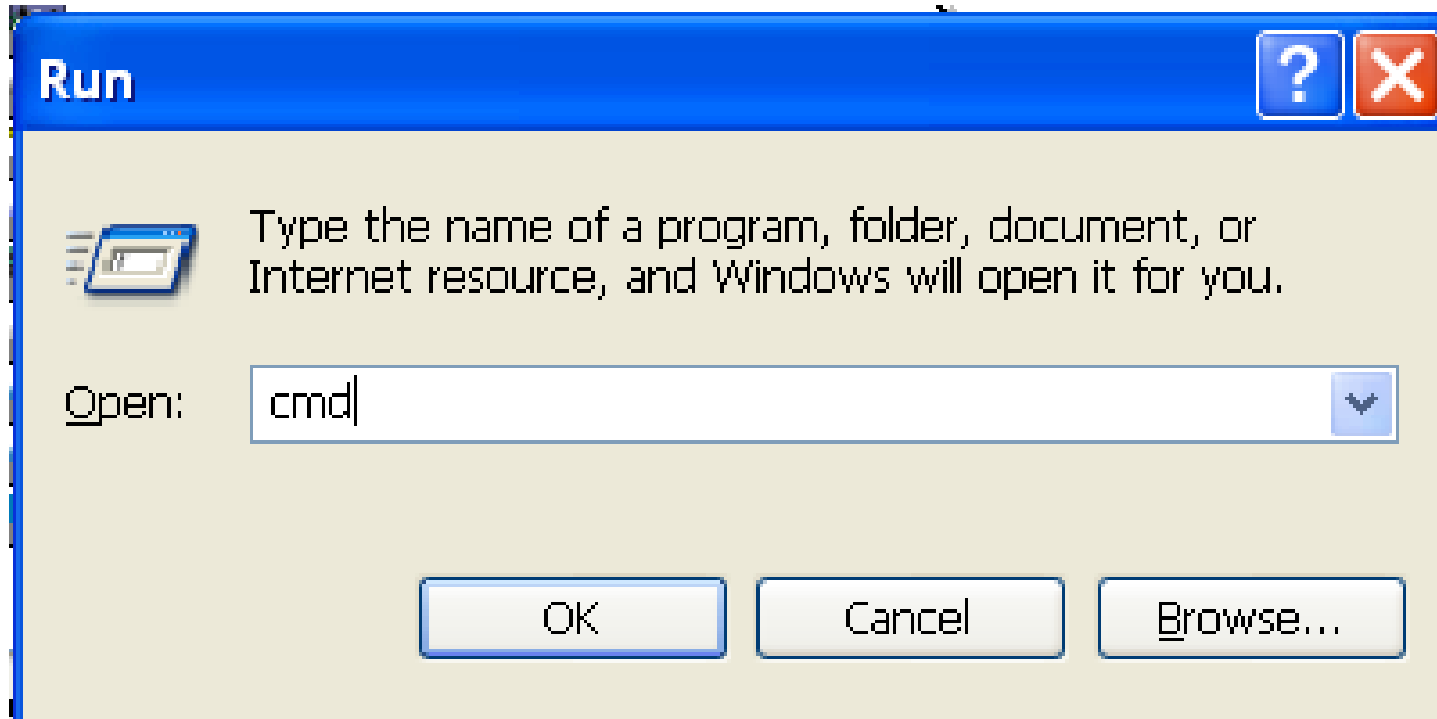
Press the **Windows Key** and
the **F** key together



F

Use DOS commands to create a file report

Press the **Windows Key** and the **R** together.
Enter **cmd** in the Run Window and click **OK**.



Consult a prophet



Cassandra's hints

File names

Two-word Canadian place names

Flin Flon

Medicine Hat

Sioux Lookout

The **pipe** argument on the filename statement creates a virtual list of data set names

```
filename indata pipe 'dir c:\data\*.dat /b';  
run;
```

```
/* Input the file names from  
the virtual list */
```

```
data _null_;
```

```
infile indata trunccover;
```

```
input prediction $ 1-256; ←
```

Maximum character
field size can be 32767

```
put _n_ prediction=;
```

```
run;
```

Log display from the **put** statement

```
1 prediction=GlanceBay.dat  
2 prediction=MooseJaw.dat  
3 prediction=SalmonArm.dat  
4 prediction=ThunderBay.dat  
5 prediction=TroisRivieres.dat
```

```
* In
data do;
  in
  in
  * Give each city a number;
  do;
    i+1;
    call symput ('full' ||trim(left(i)),
                 trim(left(fullname)));
  * R
  call symput('work' ||trim(left(i)),
              trim(left(workname)));
  call symput ('total', trim(left(i)));
end;
run;
```

Example: Glace Bay

```
%put &work1;
```

```
GlaceBay
```

```
%put &full1;
```

```
GlaceBay.dat
```

```
/* Input each city in sequence to  
learn variable names */
```

```
%macro readcities;
```

```
  %do i = 1 %to &total;
```

```
    %sales(&&work&i,&&full&i)
```

```
  %end;
```

```
%mend readcities;
```

Learning the variable names -- Consult another prophet



$$p = .0055$$

Paul's prediction of variable names

Musical instruments

Variable names

Castanet

Xylophone

```
/* Print one record to find out the  
layout of all the data files */
```

```
data glacebay;
```

```
infile "c:\glacebay.dat";
```

```
if _n_ = 1 then
```

```
do;
```

```
input @1 wholeline $256.;
```

```
put wholeline;
```


```
end;
```

```
stop;
```

```
run;
```

Examine the log for the layout

Sample line:

	1	2	2		5
1	6	0	4		6
Swan	H	6	9		941388795

<u>Variable</u>	<u>Location</u>	
Name	\$1 - 14	
Initial	\$16 - 17	
Sales1	20 - 21	Sales value for Varname1
Sales2	24 - 25	Sales value for Varname2
Varname1	\$38 - 39	Name of first variable
Varname2	\$40 - 41	Name of second variable
Invoice	\$56 - 64	

```
/* /* Convert variable names stored in columns 38 and 40 to
&sales1 and &sales2 */
% if _n_ = 1 then
fu do;
d input @38 varname1 $2. @40 varname2 $2.;
call symput("sales1",trim(left(varname1)));
call symput("sales2",trim(left(varname2)));
end;
stop;
run;
```

```
/* Use &sales1 and &sales2 in input  
statement */
```

```
data &workname;  
  infile "c:\&fullname";  
  length Location $15;  
  input Name $1-14 Initial $16-17  
        &sales1 20-21 &sales2 24-25  
  Invoice $56-64;  
  Location="&workname";  
  
run;
```

```
/* Use PROC CONTENTS to learn  
variable names */
```

```
title Variables in &workname;  
proc contents data=&workname  
out=vars&workname  
(keep=memname name) noprint;  
title Variables in &workname;  
proc sql;  
select * from vars&workname;  
quit;  
%mend sales;
```

Example of variable names

Variables in MooseJaw

Library Member Name

Variable Name

MOOSEJAW
MOOSEJAW
MOOSEJAW
MOOSEJAW
MOOSEJAW
MOOSEJAW

Initial
Invoice
Location
Name

fb
ta

Labels for musical instruments

Variable

name

ta

fb

tb

fa

Label

Pianos

Piccolos

Cellos

Harps

```
/* Concatenate the data sets */
```

```
pr %let cities_list=;  
data _null_;  
  set cities;  
  call symputx('mac',memname);  
  call execute('%let cities_list=&cities_list  
pr &mac;');  
b run;  
ru
```

```
%put &cities_list;
```

The log shows:

```
GLACEBAY MOOSEJAW SALMONARM THUNDERBAY TROISRIVIERES
```

```
/* Use set to join the files together */
```

```
data all_cities;  
    set &cities_list;  
    label fb = "Piccolos" ta = "Pianos"  
          fa = "Harps"    tb = "Cellos";  
run;
```

```
proc sort data=all_cities;  
    by Name Initial Location;  
run;
```

```
/* Print the first 15 records in the concatenated  
file */
```

```
Title Concatenated file (First 15 records);
```

```
proc sql outobs=15;
```

```
    select Name, Initial, location, ta, fb, tb, fa  
    from all_cities;
```

```
quit;
```

First 15 records in the concatenated file

Concatenated file (First 15 records)

Name	Initial	Location	Pianos	Piccolos	Cellos	Harps
Agostin	M	SalmonArm	.	.	9	7
Agostin	M	TroisRivieres	.	.	5	8
Awaw	H	GlanceBay	6	9	.	.
Awaw	H	ThunderBay	5	7	.	.
Awaw	H	TroisRivieres	.	.	5	7
Baffo	B	GlanceBay	4	10	.	.
Baffo	B	MooseJaw	8	10	.	.
Baffo	B	SalmonArm	.	.	10	11
Bonga	LS	GlanceBay	4	5	.	.
Bonga	LS	TroisRivieres	.	.	5	7
Bram	C	SalmonArm	.	.	10	11
Bram	C	ThunderBay	10	9	.	.
Brown	DE	SalmonArm	.	.	10	10
Brown	DE	ThunderBay	7	10	.	.
Brown	DE	TroisRivieres	.	.	6	6

Aggregate the files

```
/* Aggregate by name and initial */  
proc sort data=all_cities;  
  by name initial;  
run;
```

```
proc means data=all_cities noprint;  
  var ta fa fb tb;  
  by name initial;  
  output out=agg_cities (drop=_type_ _freq_)  
  sum=;  
run;
```

```
/* Print the first 15 records in the  
    aggregated file */
```

```
Title Aggregated by name (First 15 records);  
proc sql outobs=15;  
    select * from agg_cities;  
quit;
```

First 15 records in the aggregated file

Aggregated by name (First 15 records)

Name	Initial	Pianos	Harps	Piccolos	Cellos
Agostin	M	.	15	.	14
Awa	H	11	7	16	5
Baffo	B	12	11	20	10
Bong	LS	4	7	5	5
Bram	C	10	11	9	10
Brown	DE	7	16	10	16
Cammisa	A	12	.	13	.
Carreno	K	7	11	12	6
Carreno	LA	9	.	11	.
Church	R	15	11	20	10
Chyn	S	5	5	5	5
Coelho	DM	9	.	10	.
Cooper	C	9	12	12	13
Cowan	M	15	12	18	13
Currie	L	23	7	27	5 ³⁴

Thanks!

William Klein

416/482-5410

Cell 416/707-5137

E-mail: william.klein@utoronto.ca

Skype billyklein