Paper 116-2008

The Perfect Baby Name: One Dad's Quest Using SAS®

Chris Hemedinger, SAS Institute Inc., Cary, NC

ABSTRACT

If you think that baby names and data analysis have nothing to do with one another, then you haven't read *Freakonomics*. When we were expecting my third daughter and were selecting her name, we had specific criteria in mind. This paper describes how I used SAS® Enterprise Guide and data from the Social Security Administration to select a list of candidate names and how this helped us embark on the process with our eyes wide open.

INTRODUCTION

Naming your child is a big responsibility. For many people, the task involves combing through the *Big Book of Baby Names* and picking a handful that sound good, and heading into the delivery room with those candidates in mind.

But I needed to do more research to make an informed decision--and, of course, I used SAS. After all, anyone who has read *Freakonomics* knows that a name can be an indicator of so many other attributes. (The *Freakonomics* book contains a chapter that describes how the names that parents select for their children can be linked to certain life outcomes.)

BACKGROUND: THE SHAME OF BEING UNORIGINAL

It started with my first two daughters, Marguerite (with "Maggie" as a nickname) and Evelyn. We did not apply any particular methodology when selecting their names, but our friends noticed something: each name was considered "old-fashioned" and not used much anymore. At least, that was a first impression. However, as the months went by, we noticed that other families that we met had small children with the same names we had picked. We had been part of a trend, completely by accident. According to the Social Security Administration, the name "Evelyn" ranks 98th in the top 1000 girls' names of 2002. I never saw that coming.

ANALYSIS TO THE RESCUE

When we were expecting our third daughter in 2005, I decided to devise a litmus test for our preferences. That is, we know that we like "old-fashioned" names, but we don't want to take part in a "name comeback" trend. For example, my daughters attend school with a lot of girls named "Caroline" and "Sarah". This requires that we use the surname initial as a suffix when we talk about them at home. (Example conversation with our daughter: "Caroline was sent home from school with head lice today." "Oh really? Was that Caroline M. or Caroline I.?" I respond, feeling suddenly itchy.)

In plain language, our question was this: what are the names that were popular 80 to 100 years ago but aren't as popular now? Even fancy new Web tools like the Baby Name Wizard (http://babynamewizard.com/namevoyager/) don't answer that directly.

I used the following steps to find the answer.

STEP 1: DEVELOP A LIST OF ONCE-POPULAR GIRLS' NAMES

The Social Security Administration maintains a database of the historically popular baby names. The site offers different categories of queries, one of them being the "Top 1000 names by decade." You simply select a decade and the site returns a table of the top 1000 boys and top 1000 girls' names from that era with their rankings.

I used the Web site to run this query three times: once for each “old-fashioned” decade that I was interested in: 1900s, 1910s, and 1920s. The Web site output looks like this:

Most Popular Names of the 1900s

	Male			Female		
Rank	Name	Number	Percent[a]	Name	Number	Percent[b]
1	John	84,602	5.7643	Mary	161,474	5.1989
2	William	69,338	4.7243	Helen	69,408	2.2347
3	James	62,179	4.2366	Margaret	57,909	1.8645
4	George	43,590	2.9700	Anna	54,905	1.7678
5	Charles	36,193	2.4660	Ruth	51,002	1.6421

Using SAS Enterprise Guide, I brought that data into SAS. How? I could have accomplished it a few different ways, such as copying and pasting the data into Microsoft Excel and then using the Import Data task to read it in. However, I decided to write a simple SAS program to create the data set, using the data from Web site tables as input via a DATALINES statement. Here is a sample of the DATA step program:

```
DATA WORK.NAMES1900_1930;
    attrib Rank
           length=8;
    attrib BoyName
           length=$12;
    attrib BoyCount
           length=8;
    attrib GirlName
           length=$12;
    attrib GirlCount
           length=8;
    attrib Year
           length=8;
INFILE DATALINES DSD;
INPUT
    Rank
    BoyName
    BoyCount
    GirlName
    GirlCount
    Year
    ;
DATALINES4;
1,"John",25515,"Mary",24455,1900
2,"William",20247,"Helen",11468,1900
3,"James",15196,"Margaret",9638,1900
4,"George",13000,"Anna",8856,1900
5,"Joseph",12144,"Ruth",8148,1900
/* lines omitted */
;
run;
```

The result was about 3000 entries, of course (1000 per decade), but really only about 1200 unique girls’ names.

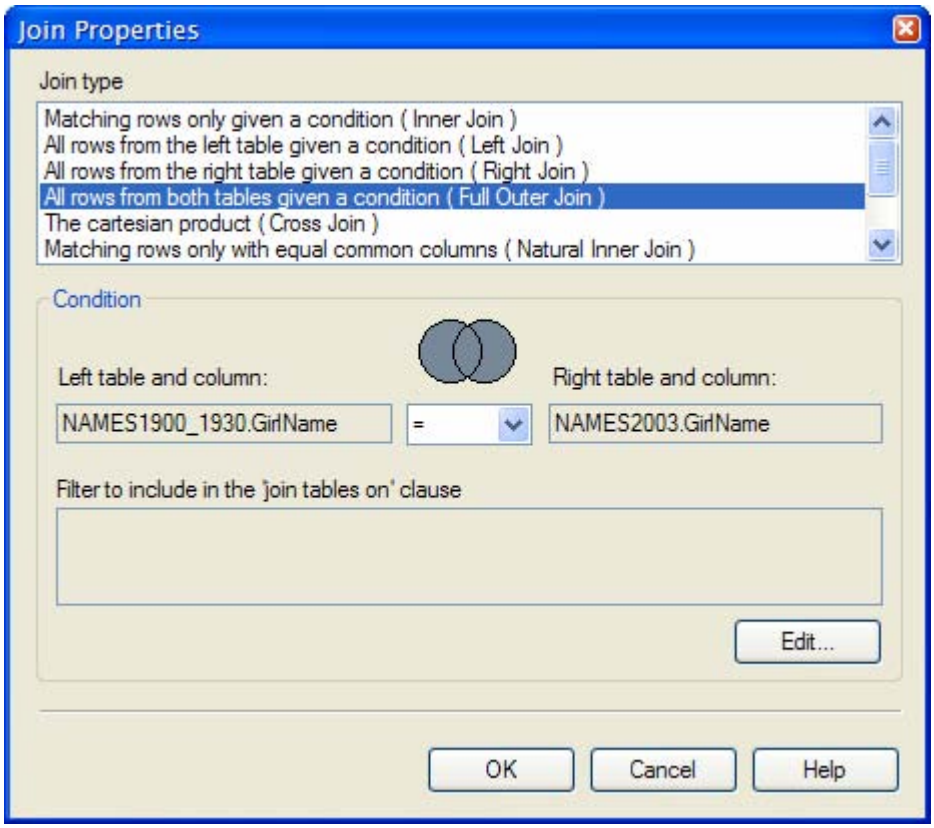
STEP 2: CREATE A LIST OF THE TOP 1000 MODERN GIRLS’ NAMES

The Social Security Administration Web site offers another query type named “Popular Names by Birth Year,” allowing you to specify the year of interest and how many names you want to see (up to 1000). Using this, I produced a list of the most popular baby names in 2003. The list has a format similar to the list in the previous step, and I imported the data into SAS using a similar technique: creating a simple DATA step with a DATALINES section for the names and numbers.

STEP 3: MERGE THE LISTS: IN WITH THE OLD AND OUT WITH THE NEW

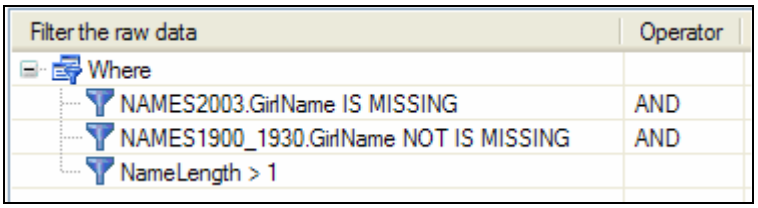
I now had two lists of names: one with names that were popular 80 to 100 years ago, and one with names that are popular now. I needed to merge these lists to keep all of the old-fashioned names **except for those** that are popular today. This process would also discard those names that are popular today but were seldom or never used 100 years ago (goodbye "Brittany" and "Brianna").

To accomplish this, I used the query tool in SAS Enterprise Guide. In query parlance, what I wanted was a full outer join of the two tables based on "girl name," where "old girl name" is MISSING (or null) in the "new girl name" list. For the full outer join, the join properties look like the following display:



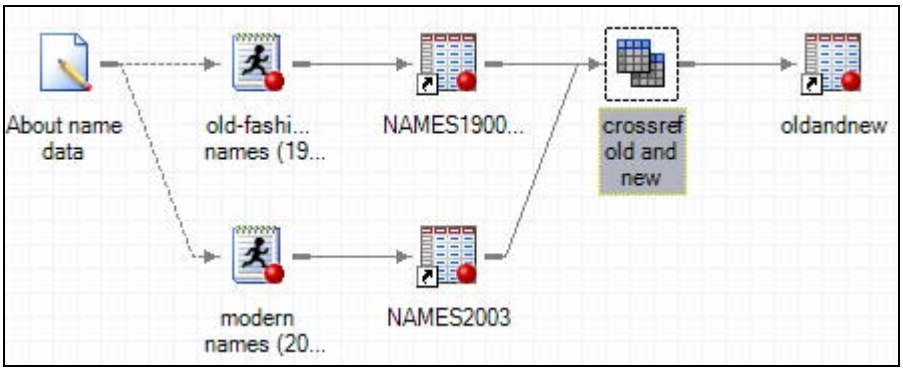
(**Note**: the screen captures in this paper were made using SAS Enterprise Guide 4.2. However, all of the features shown are available in earlier releases of SAS Enterprise Guide as well.)

The filter list in the query looks like this:



The first filter is a tricky way of saying "include the record only if there is no corresponding name in the 2003 list". The second filter ensures that I don't include records where there is no name (MISSING) in the "old-fashioned" list. And the last filter uses a calculated column that represents the name length; it's a data-cleansing step to filter out some one-character name entries that I found in the Social Security Administration data.

At this point, the process flow in my SAS Enterprise Guide project looks like the following display:



The first item in the flow is an explanatory note that I wrote to document the process. The next two items are the SAS programs that created the two lists of names. The highlighted item ("crossref old and new") is the query builder item that joins and filters the data.

The result set (the "oldandnew" table in the flow) contained over 900 names. The next step was simple: choose a name from one of these 900 names and we would be original. If other children started appearing with the same name, we would be the leader, not the follower.

MORE TO SOME DECISIONS THAN ANALYTICS

I'll be honest; there is a reason some of those old names aren't used much anymore (apologies to any Hildas out there reading this). In the end, we settled on "Gwendolyn," a name that sounds old-fashioned and yet still sweet to us. It turns out "Gwendolyn" still has some measure of popularity as a girl name. The name ranked 598th out of 1000 in the year 2003 (when 456 little girls were given that name, according to the Social Security Administration data).

But at least we walked into that name fully informed. We haven't yet met very many other little Gwendolyns, so perhaps her name will still count as one of her many outstanding, unique qualities.

LOOKING BACK: HOW WE DID

With the selection final, I still wanted to check my work. Using the data available from the Social Security Administration Web site, I was able to plot the popularity of the names we gave to our three children.

To do this, I used a third type of query available on the site: "Popularity of a Name." You can specify the name of interest and how many years of historical data you want. You also have the option to specify the sex associated with the name. I ran this query for each of our chosen names, going back 100 years for each. The report on the Web site looks something like this:

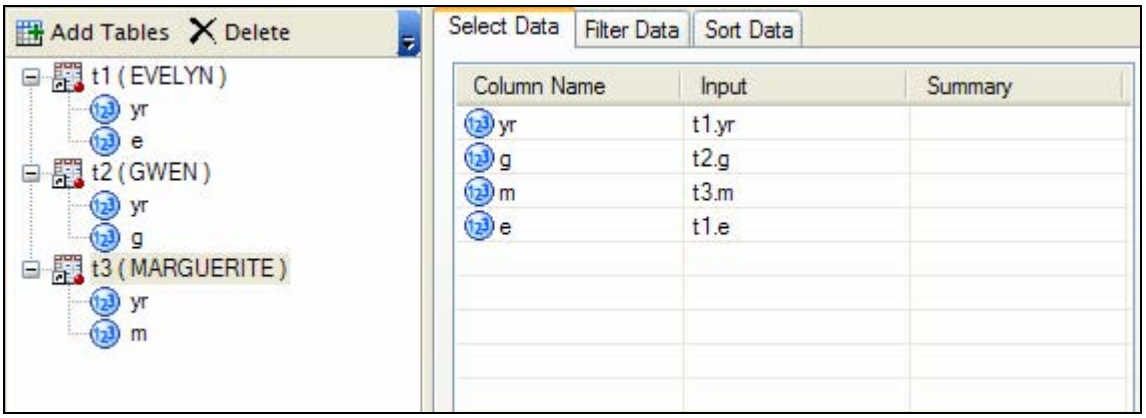
Popularity of the female name Marguerite

Year of birth	Rank
1972	966
1971	992
1970	767
1969	760

I used the output of these reports to create DATA step programs, similar to the approach I showed earlier. For example, here is the start of the program to track the trends for "Gwendolyn":

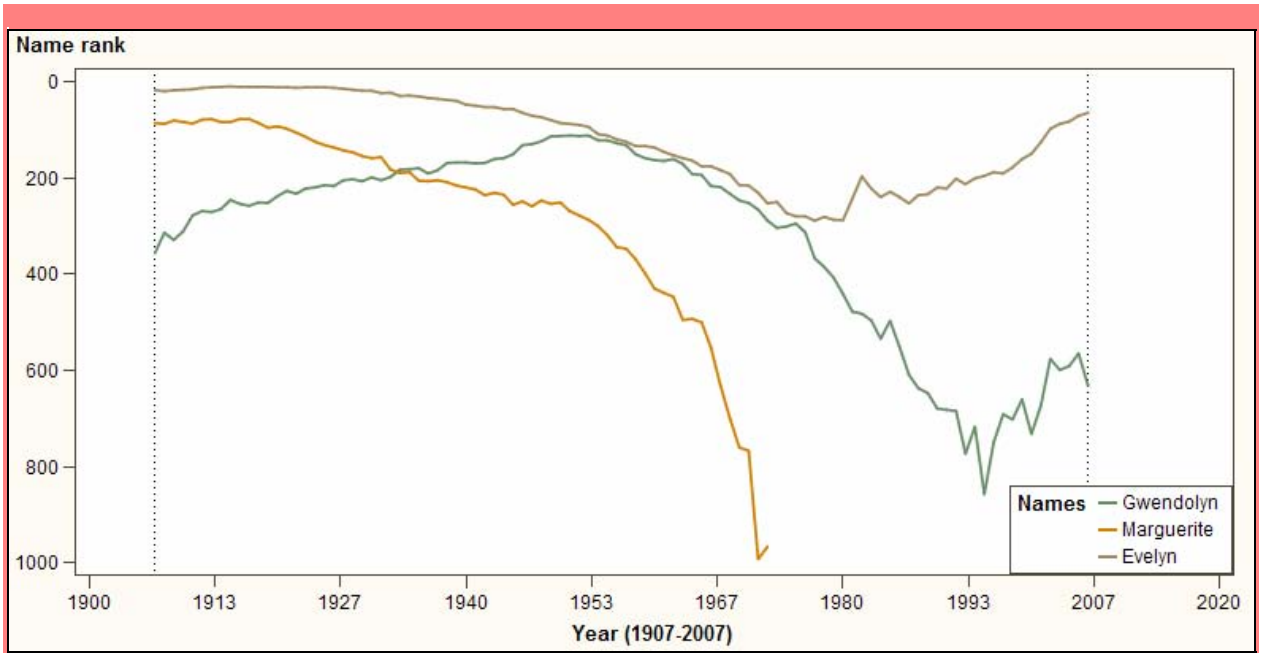
```
data gwen;
length yr 8 g 8;
infile datalines dsd dlm=' ';
input yr g;
datalines;
2006 631
2005 565
2004 591
2003 599
2002 576
/* lines omitted */
run;
```

These three DATA step programs yield three data sets, which I then joined together using the query builder tool. I used a full join on each table combination to ensure that there was a record for each year, even if one or more of the names did not appear in the ranking for a given year. Here is a display of the query selection:



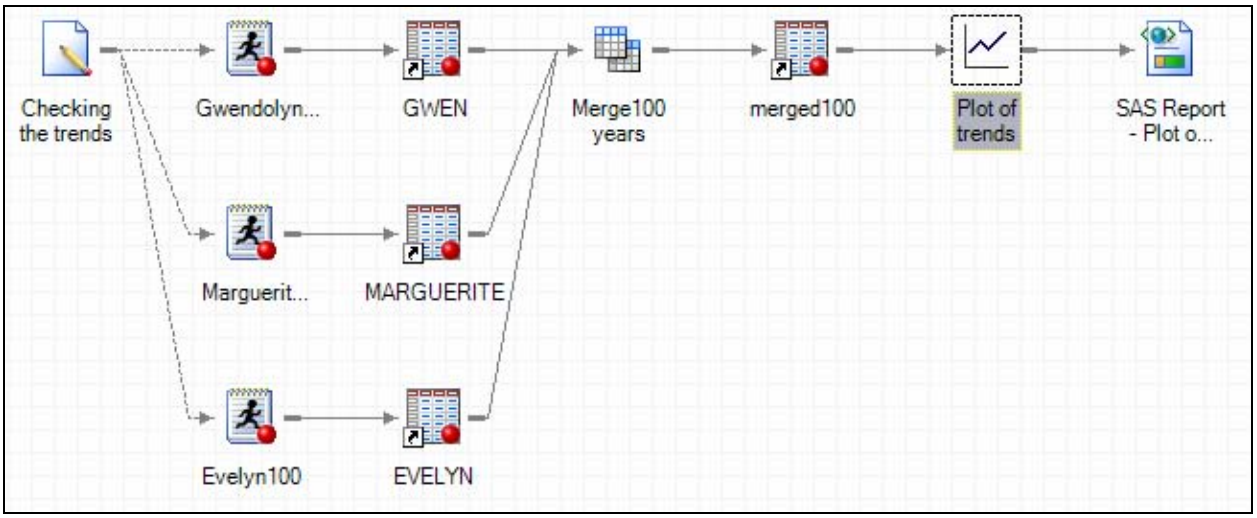
The result of the query is a merged data set with four columns: one for the year and one for each name and its ranking in that year. This data is now ripe for plotting using the Line Plot task in SAS Enterprise Guide. The Line Plot task offers several types of plots. I selected the "Multiple vertical column line plots using overlay" option, which lets me put the year on the horizontal axis and the rankings for each of the three names on the vertical axis.

I used a few of the many options in the Line Plot task to add vertical reference lines and customize the tick marks on the axis and the legend. Here is the result:



As you can see, we bucked the trend with Marguerite (named after my maternal grandmother). Her name (shown by the yellow line) drops out of the top 1000 list after 1972. But it appears that Evelyn (the brown line along the top) can expect to encounter many peers who share her first name. With Gwendolyn (the green line), we ended up somewhere in the middle. **Note**: The plot above shows *ranking* trends only, not *frequency of use*. We cannot draw conclusions about how often a name is used in relation to other names; we can infer its popularity based only on its placement in the top 1000 names.

Here is a snapshot of the process flow that produced the ranking data, combined it into a single data set, and then created the plot:



CONCLUSION

As an avid user of SAS and a developer on SAS Enterprise Guide, I see data everywhere, even in baby names. While the use of SAS for “baby name optimization” might not be a valuable commercial application, the process did help provide our family with some interesting insight to make an informed decision.

REFERENCES

Hemedinger, Chris. “The Perfect Baby Name: One Dad's Quest Using SAS” (with links to the SAS Enterprise Guide project files used in this paper). Available http://blogs.sas.com/sasdummy.

iVillage. The Baby Name Wizard (an interactive Web site that shows the frequency of use for baby names in the United States). Available http://babynamewizard.com/namevoyager/.

Social Security Administration. Popular Baby Names. Available http://www.ssa.gov/OACT/babynames/.

ACKNOWLEDGEMENTS

Thanks to my colleagues who reviewed this paper and offered feedback: John Alspaugh, Mike Porter, Bill Sawyer, and Eileen Taht.

RECOMMENDED READING

Levitt, Steven D., and Stephen J. Dubner. 2005. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything.* New York: HarperCollins.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chris Hemedinger
SAS Campus Dr.
Cary, NC 27513
E-mail: chris.hemedinger@sas.com
Web: http://blogs.sas.com/sasdummy