Paper 3738-2019

Improving Survey Data Quality and Use with SAS® Data Management Studio and SAS® Visual Analytics

Ankita Kalita, Bonnie Chapman, Institute for Veterans and Military Families, Syracuse University

ABSTRACT

The Institute for Veterans and Military Families (IVMF) offers many nationally run programs, and the survey data we collect across our entrepreneurship programming portfolio captures business outcomes of participants. The cleaning methods in SAS® Data Management Studio, which include a number of sequel executes and data jobs with expression, standardization, concatenation, data validation, clustering, and surviving record nodes, are discussed. With the cleaned data, dashboards were built in SAS® Visual Analytics to communicate program outcomes. The presentation walks through the rationale behind the evaluation and analysis, and how to conduct each step from cleaning the raw data through to its presentation in SAS Visual Analytics. It also details the way in which IVMF uses graduate student talent, the hallmark of our success at IVMF, higher education's first interdisciplinary academic institute that is focused on advancing the lives of the nation's military veterans and their families. As a nonprofit situated on the Syracuse University campus, the IVMF is uniquely positioned to optimize students across 13 schools and colleges, while providing them invaluable real-life experience.

INTRODUCTION

The Institute for Veterans and Military Families (IVMF), located on the Syracuse University campus, has three national facing program portfolio areas in addition to its research and evaluation efforts. One portfolio is a suite of Entrepreneurship programs where business sustainability and growth are markers of success. An annual survey was created that asks past program participants to provide outcome data on each of the businesses they owned in the previous year.

The annual survey was created in Qualtrics, a survey administration tool, and the output data is a wide flat file. The Qualtrics output data structure was completely incompatible to support the reporting needs. The desired data structure necessitated extensive data transformation and cleaning, so much so that SAS Data Management Studio was chosen as the tool for the job.

The following paper describes the process we used to clean our data in SAS Data Management Studio and the formatting of the data file we needed for use in SAS Visual Analytics.

Figure 1. Data Management Studio Nodes used in this example are: Data Source, Expression, Standardization, Concatenate, Clustering, Surviving Record, Field Layout, and Data Target.

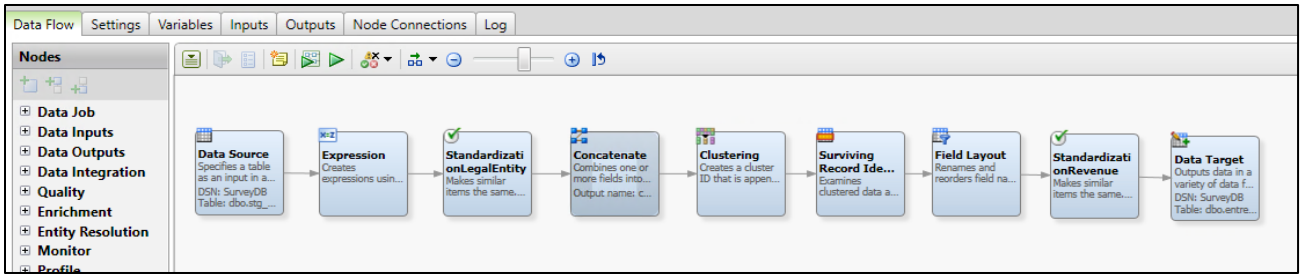

Figure 1. Data Management Studio Nodes used

DATA STRUCTURE AND STORAGE

The source data is a wide, 901 column, csv file that originates from Qualtrics and is imported to a SQL server database. An ODBC connection was created from Data Management Studio to the SQL server database to access the raw data, clean the data, and insert the clean data into a new database table. SAS Visual Analytics accesses the cleaned data by connecting to the SQL Server database using a defined ODBC connection.

THE ORIGINATING DATA STRUCTURE

Annually, a survey is sent to former Entrepreneurship program participants and they are asked to provide information about their business(es). A survey respondent may take the survey any number of years and they are primarily identified by their email address. Each year a participant takes the survey is a new record. Sometimes, a participant's delay in completing the survey resulted in the creation of additional record.

So, a participant may have more than one record per year. Each record contains all of the participant's business information and the participant may own more than one business.

Table 1 following provides a list of the type of fields in the originating data file where the fields described below are columns.

Example Fields	Comments
Date/time survey started, date/time survey finished	Meta data is collected on the survey and respondents have unique survey response IDs
Participant name, email, program type	Basic demographics are collected on the person
Business 1: name, year business starts, states where business is registered, number of employees	Over a dozen questions are asked related to business 1
Businesses 2-5: name, year business starts, states where business is registered, number of employees	The same questions are asked for each of up to 5 businesses a participant might own
Reasons for not starting a business	There is a section of questions related to why a participant did not start a business
Other research related questions	Some years this survey is sent, it includes extra research questions

Table 1. Example of original data fields and descriptions

DATA TRANSFORMATION: SQL EXECUTE NODE

We first transposed the data using the SQL Execute Node in the process flow. This node uses the ODBC connection we created and it runs the SQL queries. In the Code Editor terminal, we wrote the actual code to transform the data.

Figure 2. SQL Execute Nodes and a sample from Code Editor terminal used to transpose each of the five sets of business questions.

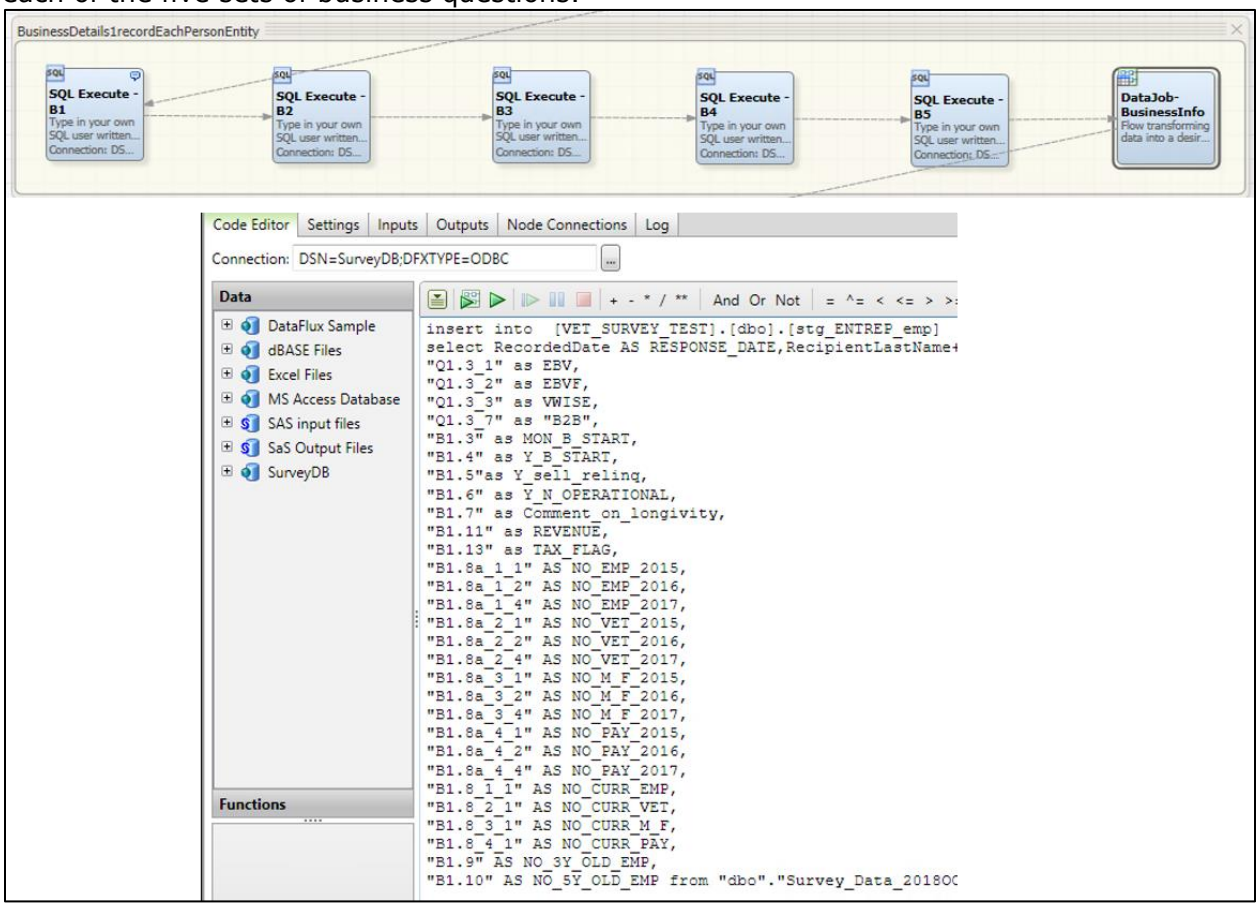

Figure 2. SQL Execute Nodes and Code Editor

We created five SQL Execute Nodes to transpose data for all five business question groupings and inserted the data into the single table. The SQL Execute Node also has features which allow it to run SQL statements before and after running the main SQL statements. We also used the SQL Execute Node to delete the data of the stage tables we used when we created the transposed data.

The column names in the source dataset are specific to each of the 5 business groupings. All columns pertaining to the first business start with "B1" and columns pertaining to the second business start with "B2" and so on. The SQL Execute Node is also used to rename the columns to single column name while transforming the data.

Table 2example of the original data structure before it was transposed.

RECIPIENT EMAIL	Q1.5	B1.1A	B1.2 ~ B1.15	B2.1A	B2.2 ~ B2.15	B3.1A	B3.2 ~ B3.15	B4.1A	B4.2 ~ B4.15	B5.1A	B5.2 ~ B5.15
Example @gmail.com	3 businesses	Name Business 1	~ ~ ~ ~	Name Business 2	~ ~ ~ ~	Name Business 3	~ ~ ~ ~	NULL	~ ~ ~ ~	NULL	~ ~ ~ ~

Table 2. Original data structure

The resulting table contains duplicates and uncleaned data, which we processed in the next steps.

Table 3 this:

RECIPIENT EMAIL	Legal Entity	Business Details(COL 3 ~ ~COL N)
example@gmail.com	Name Business1	~ ~ ~ ~ ~ ~ ~ ~ ~
example@gmail.com	Name Business2	~ ~ ~ ~ ~ ~ ~ ~ ~
example@gmail.com	Name Business3	~ ~ ~ ~ ~ ~ ~ ~ ~

Table 3. New data structure

DATA CLEANING: STANDARDIZATION SCHEME BUILDER

Once the data was transposed, we cleaned the data using a Data Job in the process flow.

The business name, or legal entity as it is described in the survey, is a free text entry field. Unfortunately, this name is needed as a unique identifier for the business. The survey was designed so that participants would see the name that they previously entered; however, many still wrote a name in the field that varied ever so slightly from the original name they entered. The Schema Builder feature helped us to address this challenge. We first created a profile on the legal entity field and then created a Standardization Scheme, which clustered the legal entity names.

Figure 3. Example using the Standardization Scheme

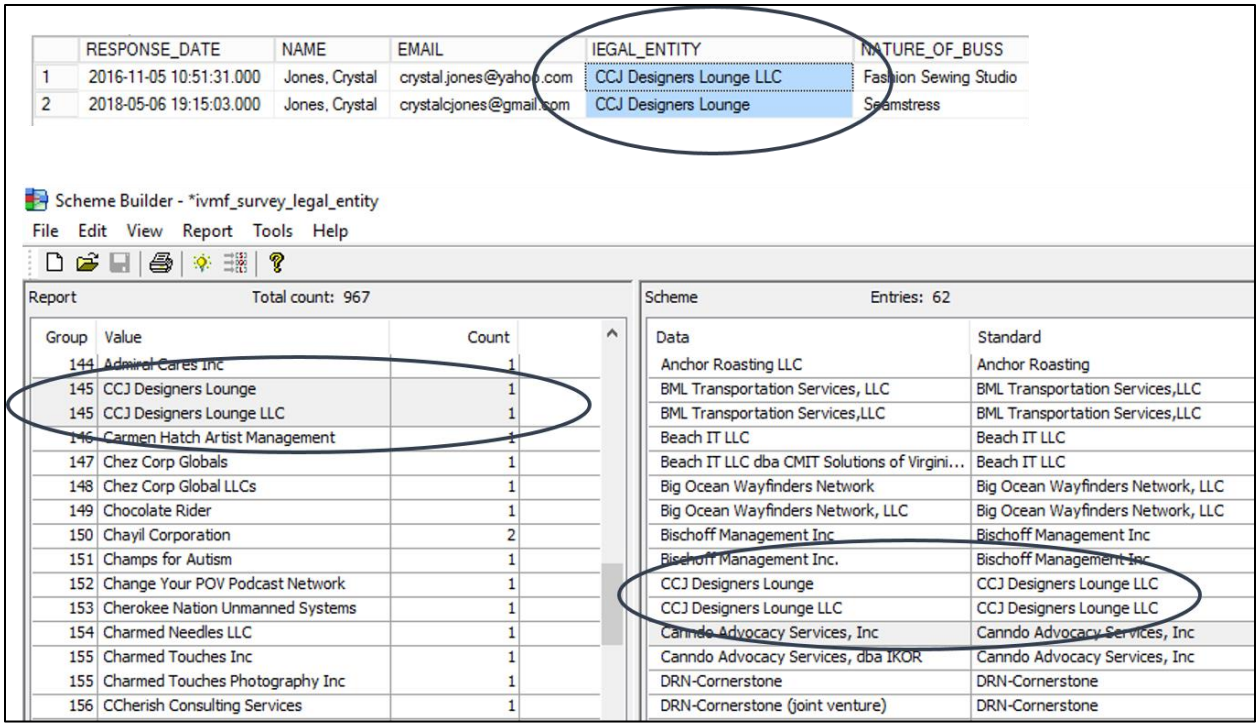

Figure 3. Example using the Standardization Scheme

DATA CLEANING: CONCATENATE NODE

We used the Concatenate Node to uniquely identify the business name by person and year. The Concatenate Node concatenates on email, business name, and year the survey was recorded. The new resulting field, concat_id, is used in the Clustering Node later on for further data cleaning. Specifically, the Concatenate Node does the following:

Concat_id="<email>_< year the survey was recorded >_<business/legal entity>"

DATA CLEANING: CLUSTERING

Due to inconsistencies in data entry, some records are incomplete or multiple records per person for one year may exist. The table below shows how even within a single year there could be two records with one partially completed. Table 4. Data prior to clustering

Concat_id	Business Detail COL 3	Business Detail COL 4	Business Detail COL 5	Business Details(COL 6 ~ ~COL N)
example@gmail.com_2018_NameBusiness1	Values	Values	Values	~ ~ ~ ~ ~ ~ ~ ~ ~
example@gmail.com_2018_NameBusiness1	NULL	NULL	NULL	~ ~ ~ ~ ~ ~ ~ ~ ~

Table 4. Data prior to clustering

In the Clustering Node we selected the field we created in the previous step, Concat_id, to be the basis for the clustering. This clustering step creates an ID for all clusters, which we use in the next step. Figure 4. Clustering step

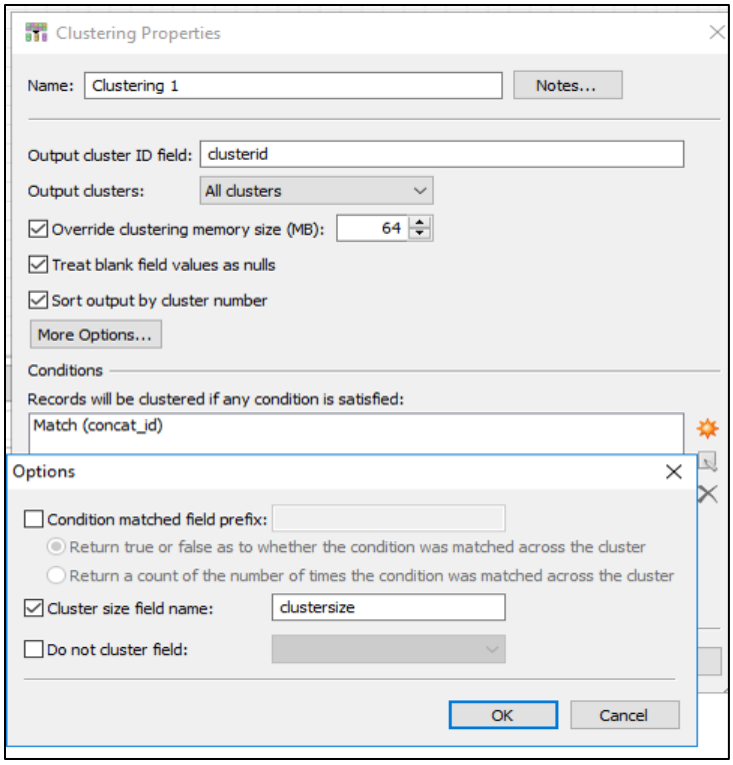

Figure 4. Clustering step

DATA CLEANING: SURVIVING RECORD IDENTIFICATION NODE

In the final major data cleaning step, we needed to choose a surviving record from the clusters. The Surviving Record Identification Node uses the cluster ID we created in the previous step. This step allowed us to take values from any field from the records within a single cluster ID and combine them all in one surviving record. Figure 5. Surviving record identification

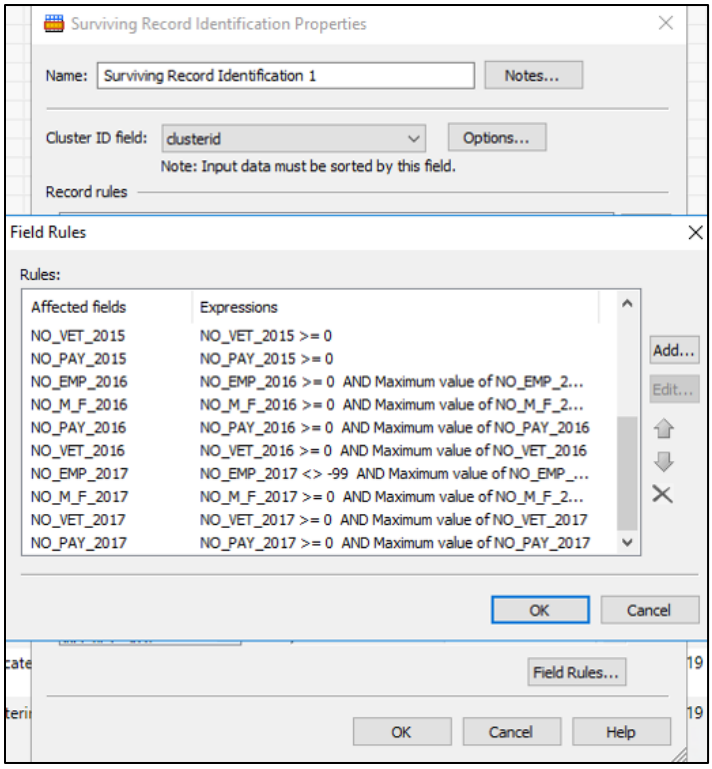

Figure 5. Surviving record identification

DATA VISUALIZATION

The cleaned dataset was analyzed to address specific outcome reporting needs of the entrepreneurship programs. Data visualizations were created in SAS Visual Analytics to show results by sub program, by businesses, by participant, over time, and by geography.

CONCLUSION

Survey data from large surveys is rarely, if ever, clean and structured in a way that is suitable for analysis. Data cleaning is needed to account for anomalies in survey taker behavior, contradictory data entry on behalf of the survey taker, or from the use of unstructured or free text entry fields, among many other reasons. SAS Data Management Studio provides an efficient and relatively simple way to perform complex cleaning steps. Having a wealth of properly cleaned and available survey data is an enviable position to be in, especially when the results speak directly to the impact of your programs.