

Show Me the Money! Text Analytics for Decision-Making in Government Spending

Tom Sabo, SAS Institute Inc., Arlington, VA

ABSTRACT

Understanding organizational trends in spending can help overseeing government agencies make appropriate modifications in spending to best serve the organization and the citizenry. However, given millions of line items for organizations annually, including free-form text, it is unrealistic for these overseeing agencies to succeed by using only a manual approach to this textual data. Using a publicly available data set, this paper explores how business users can apply text analytics using SAS® Contextual Analysis to assess trends in spending for particular agencies, apply subject matter expertise to refine these trends into a taxonomy, and ultimately, categorize the spending for organizations in a flexible, user-friendly manner. SAS® Visual Analytics enables dynamic exploration, including modeling results from SAS® Visual Statistics, in order to assess areas of potentially extraneous spending, providing actionable information to the decision makers.

INTRODUCTION

Government organizations are often finding their budgets tightened while being asked to deliver an existing or greater amount of capability. As such, these organizations might be seeking to make modifications to their spending while continuing to serve the citizenry in an effective way. They would greatly benefit from a capability to analyze historical spending data and provide actionable information to cut or modify spending. Such capability is of import to an organization's CFO/CTO/CIO office as well as any existing vendor management office. Overseeing agencies such as the Government Accountability Office and the General Services Administration can implement such practices government-wide.

However, these organizations are often limited to basic reporting on the data due to a lack of investment in analytics technology, including text analytics. While information about organization spending information is available in structured text and available to reporting solutions, much information is available in unstructured data, which can be untapped via text analytics. Without such analytics capability, organizations must resort to manual analysis of unstructured data, which has its limitations. First, it primarily produces subjective results, which cannot help you put actual data points in the proper context. Second, manual analysis takes an extraordinary amount of time and resources, which the organization might or might not have.¹ In addition, other structured data analysis capabilities can generate actionable insight that goes far beyond basic reporting.

To address these shortcomings, SAS proposes a combination of technology including SAS® Contextual Analytics, SAS® Visual Analytics, and SAS® Visual Statistics to enable structured and unstructured data analytics in order to understand trends in spending for a government agency and to extract key information for downstream reporting. This process builds upon a previously presented paper at SAS Global Forum 2014 to define a research analytics framework for government spending.²

The five-step process for generating and using the framework for research analytics is as follows:

1. **Data acquisition and preparation for text analytics:** Data is acquired for our example through web interfaces and is converted into a SAS data set using SAS® Enterprise Guide®.
2. **Text analytics:** We will use SAS® Contextual Analysis for ad hoc exploration and rule-building techniques on categorical data.
3. **Data preparation for visual analysis:** Both text analytics approaches generate data sets that are prepared for ingestion into SAS Visual Analytics using SAS Enterprise Guide.
4. **Ad hoc exploration and modeling:** This is accomplished with SAS Visual Analytics and SAS Visual Statistics.

5. **Interactive report generation and use:** This is also accomplished with SAS Visual Analytics.

RESEARCH ANALYTICS FRAMEWORK FOR GOVERNMENT SPENDING AT THE US DEPARTMENT OF TRANSPORTATION

For the purposes of illustrating this framework, we chose to obtain data from USAspending.gov,³ which was made available to provide the public with information about how their tax dollars are spent. The data obtained from this website includes millions of individual transactions for public-facing government agencies, including the Department of Defense, and each transaction includes dozens of data fields, such as the contracting agencies, spending codes, vendor information, and obligated amount. We chose to focus on an individual agency, namely the US Department of Transportation (henceforth referred to as DOT), but this process could be applied across many agencies, and we encourage all text analytics users to continue this work applying new methods against this data. Finally, we chose to focus on software spend and the associated infrastructure (hardware, training, and services) to support it.

These are some of the questions we can ask of the spending data using this framework:

- What vendors do organizations use? How can this be portrayed over time? One challenge with the data is that the structured data regarding vendors does not include the specific technology sold. For example, Oracle is resold across DOT from several vendors, but there are no structured data elements designating "Oracle technology." Thus, we need an alternative way to identify software technology and solutions to determine trends associated with it.
- What large projects or initiatives are being supported? Again, this is not easily ascertained from the structured data.
- How can we assess data quality of structured data using information from unstructured data? Fixing data quality problems will help in understanding spending in the future.
- Where is there extensive software spending? What software? What vendors? What projects? This will help DOT identify areas of potential savings.

DATA ACQUISITION AND PREPARATION FOR TEXT ANALYTICS

We obtained the data using the interface at <http://www.usaspending.gov/data>. We obtained 150,000 rows of data associated with the DOT spanning back to 2006 as an .xls data file. We imported this data into SAS Enterprise Guide, and executed the following code to flag documents with descriptions that contained the word "software":

```
libname mylib "C:\SAS\Data";  
  
data mylib.dot_sw_spend;  
set mylib.dot_spend;  
if prxmatch("/software/", lowercase(contractdescription)) >= 1 then  
sw_flag = 1; else sw_flag = 0;  
run;
```

This was immediately followed up with a SAS Enterprise Guide query builder to filter out all non-software oriented observations to give us a SAS data set with 85 columns and 2803 rows. The columnar data was mostly structured, but included a contract description field, which ranged in content from one line to a paragraph of text.

TEXT ANALYTICS—CONTEXTUAL ANALYSIS

After registering the SAS data set using SAS® Management Console, we were able to select and load the data set within the SAS Contextual Analysis interface. This enabled us to select the contract description as the field to perform text analytics against, as well as specify the principalnaicscode as a category variable. (See Figure 1 below.) The principalnaicscode is taken from the North American Industry

Classification System (NAICS). The purpose of defining a category variable is to set up groups of documents to generate textual rules against. By using principalnaicscode as a category variable, where this code designates a certain division of government spending, this enables the SAS Contextual Analysis engine to generate text rules from the contract description that differentiate spending for these different spending codes. We will explore this further shortly.

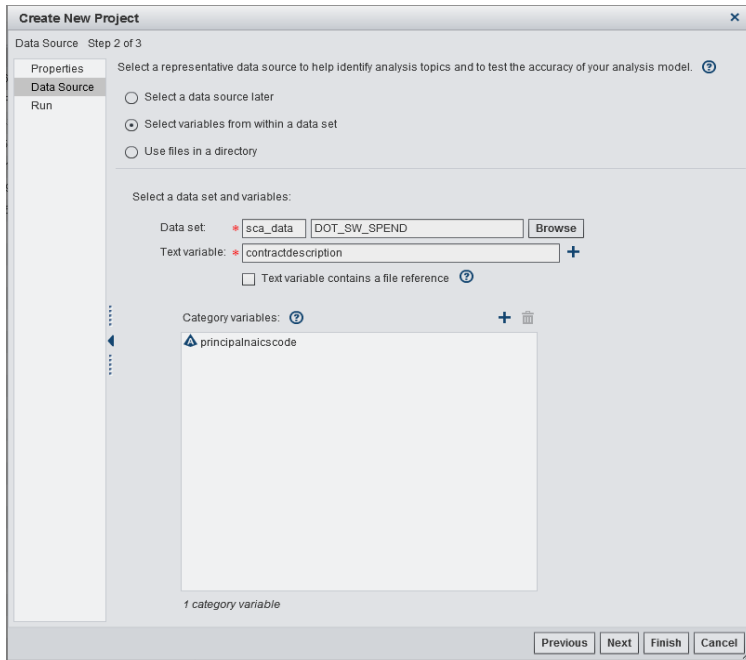


Figure 1. Defining a New Project in SAS Contextual Analysis Including a Category Variable

SAS Contextual Analysis produces a term-by-frequency matrix (see Figure 2 below), and also enables the user to explore a term map of words and phrases that co-occur. In the context of this data set, it provides an excellent opportunity to explore various software solutions and vendors to learn more about the context for how each of these software providers is being used. Oracle is one of the terms that commonly occurs in the software-oriented spending data set, and the term map for Oracle shown in Figure 3 indicates that Oracle is largely being used for its business intelligence software as part of the PHMSA hazmat intelligence portal.

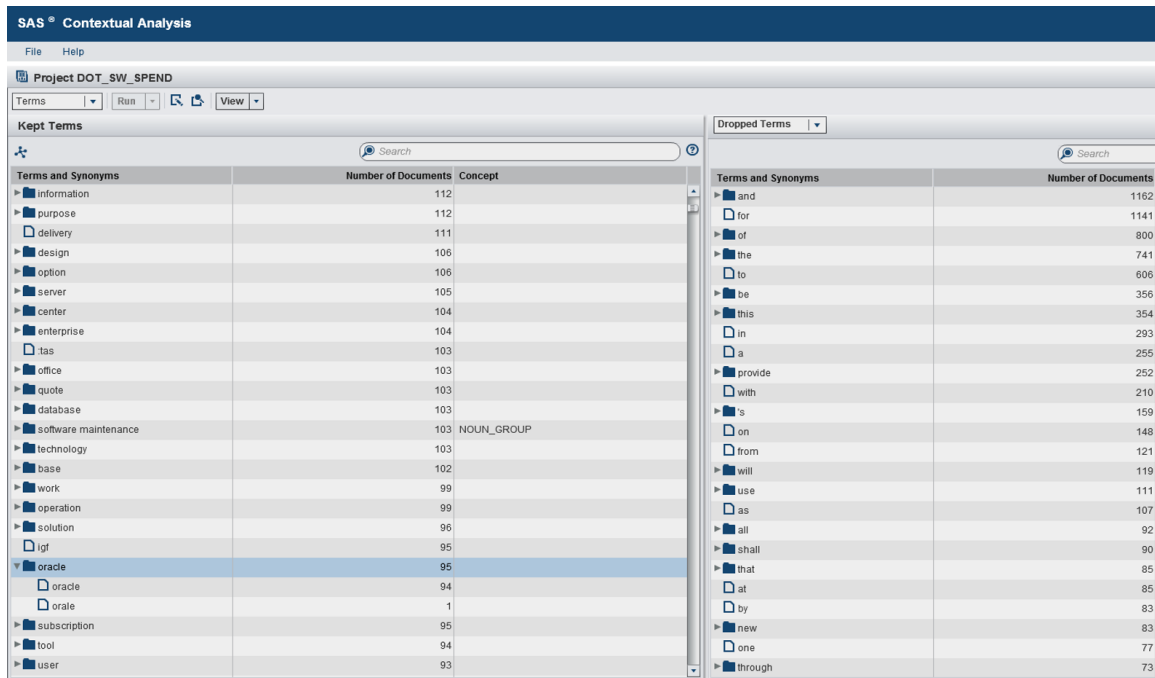


Figure 2: SAS Contextual Analysis Term-by-Frequency Matrix for DOT Spending

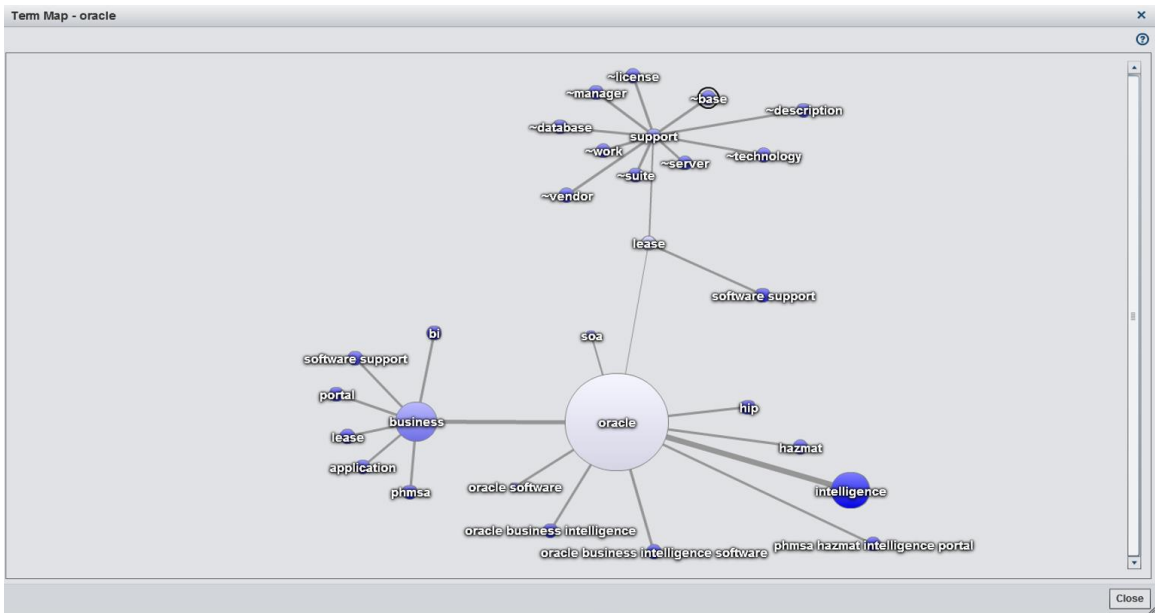


Figure 3: SAS Contextual Analysis Term Map Indicating Co-occurrence with Oracle

While SAS Contextual Analysis includes other capabilities in term exploration and topic generation, we will focus here on its capability to generate textual rules by categorical data. For a further study of terms and topic exploration in the context of a research-oriented data set, please see SAS Global Forum paper 061-2014, "Uncovering Trends in Research using Text Analytics with Examples from Nanotechnology and Aerospace Engineering."²

Oracle, SAS, and IBM are likely software solutions that are part of the DOT spending data set. How do we determine what other software is in use at the DOT? This is where the textual rules we are able to generate from the contract description against the principalnaicscode category come into play. As it turns out against this data, most of the differentiating rules generated for each NAICS code designate various software products or solutions that the DOT has acquired, is using, or is paying maintenance against. As

mentioned earlier, this is not necessarily information that is captured for each vendor in the structured data. So, this automated method of applying rule-building analytics against the NAICS code is an excellent way to determine all of the various software products and solutions in use at the DOT.

As indicated in Figure 4 below, all SAS software falls under a code for Electronic Computer Manufacturers, along with Dell, Otrak, NetIQ, HP9000, Visio, and other storage area network (SAN) services. Could this be an error in coding since SAS is not a computer manufacturer? While we do not explore it further within this paper, SAS Visual Analytics can be used to explore how contracts related to software applications span multiple spending codes, and perform a quality check on these.

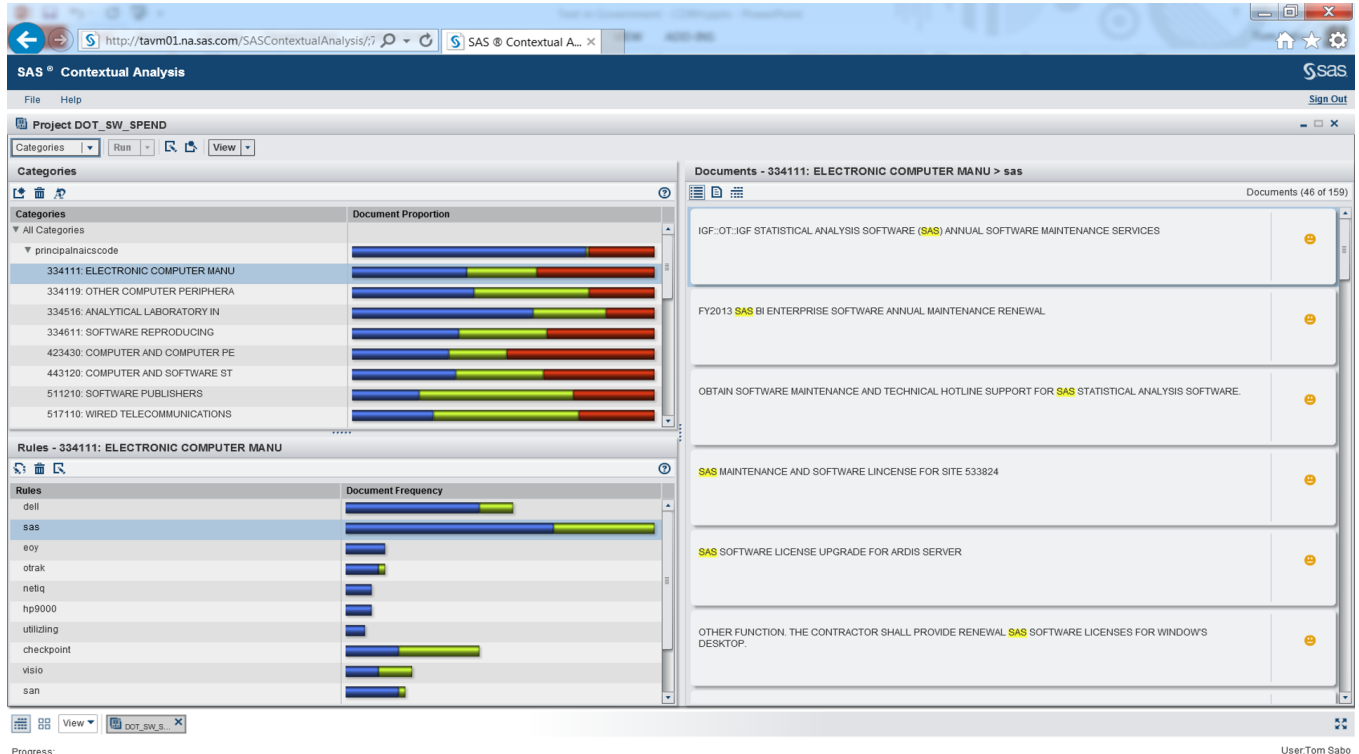


Figure 4: SAS Contextual Analysis Depicting Auto-Generated Rules for Electronic Computer Manufacturers Spending

In comparison, the code for Computer and Software Stores encompasses a significant amount of IBM and Oracle software, but not all of it. In the Figure 5 below, IBM and Oracle software specifically designated to this code is depicted in blue, while the yellow portion of the lines for IBM and Oracle designate use of these software providers in a different NAICS code context. Note some of the other software solutions in use, including “Madymo” for analyzing and optimizing occupant safety designs, “Slope Stability” software for computing the factor of safety for earth and rock slopes, and “Prism” software for automated document acquisition and management. Note how the rule generated for Prism designates the term “Prism” specifically in the context where the term “tas” does not occur. SAS is able to determine that when those two terms co-occur, it is generally in a different NAICS spending context.

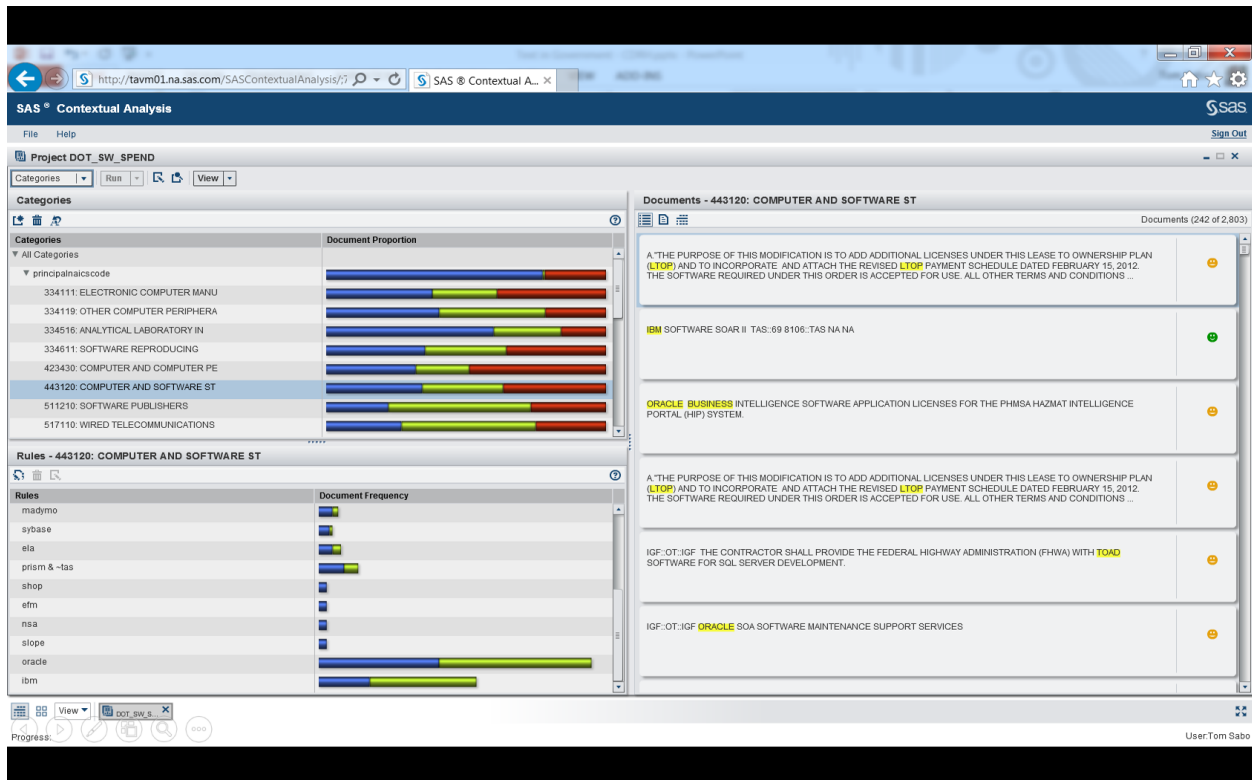


Figure 5: SAS Contextual Analysis Depicting Auto-Generated Rules for Computer and Software Spend

DATA PREPARATION FOR VISUAL ANALYSIS

In order to prepare the data for visual analysis, we are required to link a couple of the tables that SAS Contextual Analysis produces, and we used SAS Enterprise Guide to accomplish this as depicted in Figure 6 below. Doc_rule_ds contains all instances where a rule fired for a particular NAICS code, such as "oracle" or "sybase." Doc_category_rule includes all the original data we analyzed, including both structured data and free-form text. Both tables include a document ID, which is the primary key for linking information from these two tables. For demonstration purposes, we chose to import a subset of the software solution/vendors that were identified as category rules from SAS Contextual Analysis. This exercise produced a data set with all the original structured data plus a number of binary variables designating whether the spending line item for DOT dealt with a specific software solution or vendor, such as SAS, Oracle, or IBM.

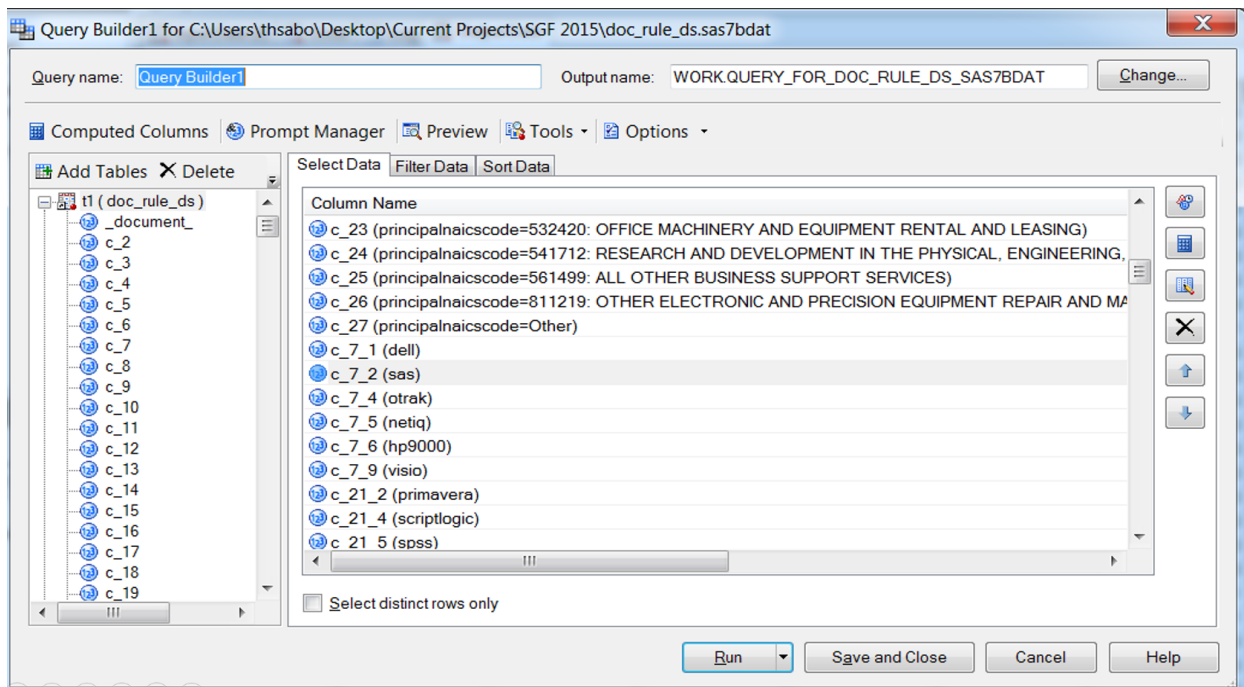


Figure 6: SAS Enterprise Guide Depicting Data Preparation for Visual Analysis and Reporting

AD HOC EXPLORATION AND MODELING

We load the data into SAS Visual Analytics for exploration, modeling, and reporting. Applying a correlation analysis, as depicted in Figure 7 below, highlights linkages between the various products and vendors, as well as certain offices of the DOT. For example, there is a correlation between HP products and annual revenue, indicating a greater amount of spending on these products, as well as SAS and Volpe, indicating prevalent SAS usage at that particular DOT agency. Volpe is a federal agency under the DOT responsible for improving transportation by anticipating and addressing emerging issues and advancing technical, operational, and institutional innovations across all modes⁴, so it is valuable to note the prevalent SAS usage at that agency versus other analytics vendors. While the correlation between SAS and Volpe is designated “weak” in the Visual Analytics correlation diagram, it stands out from the other correlations in the diagram, and even a “weak” correlation denotes some substantial overlap when considering correlation between text analytics categorical results.

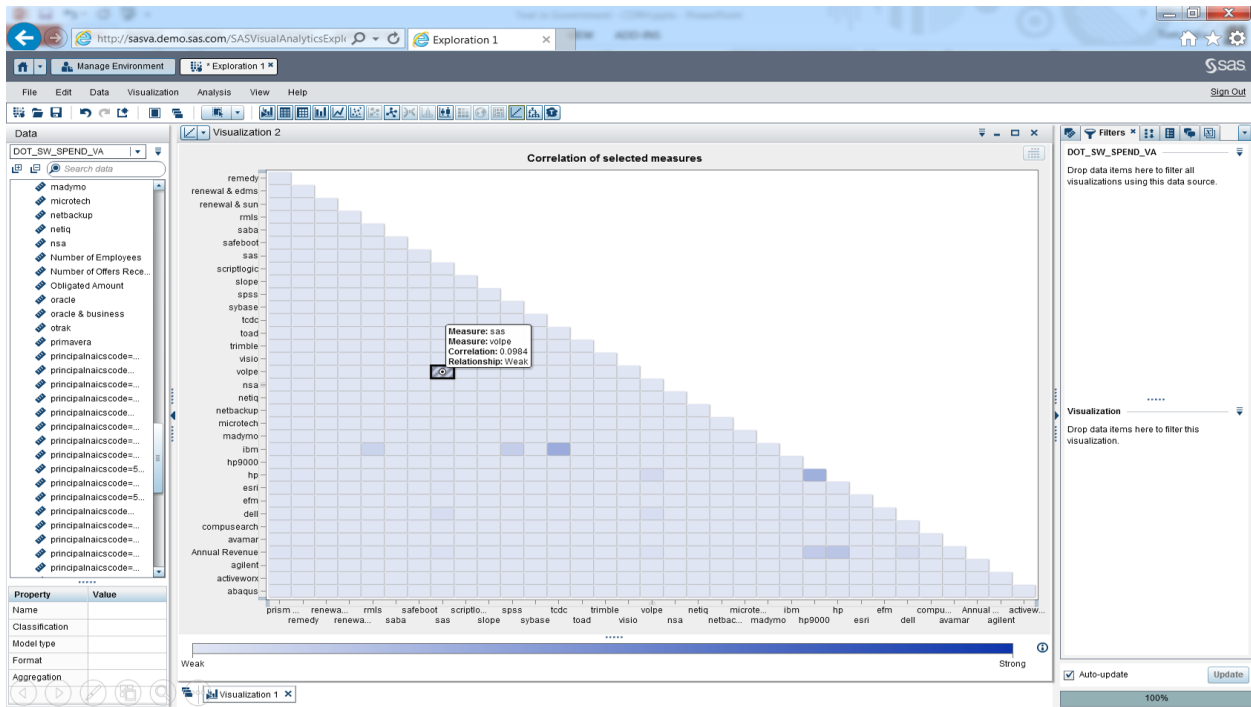


Figure 7: Correlation Analysis of Text Generated Categories Using SAS Visual Analytics

Tile charts against hierarchies in SAS Visual Analytics are also useful for exploring vendors and the software that they offer. In the tile chart depicted in Figure 8, we explore a particular NAICS code for computer and software stores. The color coding in the diagram is indicative of which vendors offer Oracle solutions. Oracle solutions are primarily provided by DLT Solutions LLC. However, they are also registered under DLT Solutions Inc. Are these two separate organizations or is this a data quality issue? In addition, there are other vendors that provide Oracle software and services, such as Mythics Inc. The DOT can review the relationships with the multiple vendors for Oracle services/software and determine if there are cost savings possible with consolidation. The second tile chart depicted in Figure 9 is a further drilldown into DLT Solutions LLC to see, at the line-item level, the specific descriptions of all DOT spending line-items associated with DLT Solutions, and which ones correlate with Oracle software.

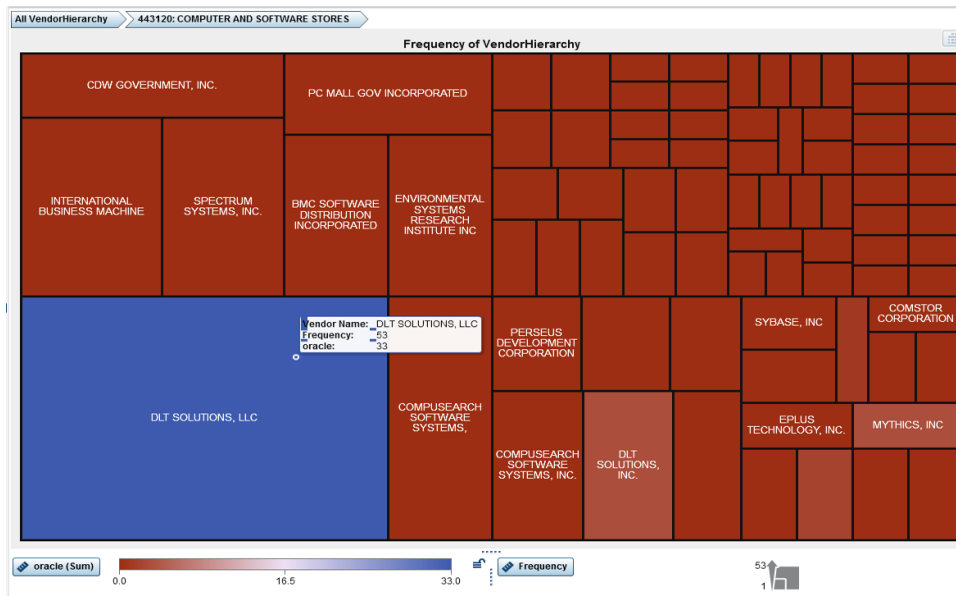


Figure 8: Tile Chart of Vendor Hierarchy by Oracle Spending in SAS Visual Analytics

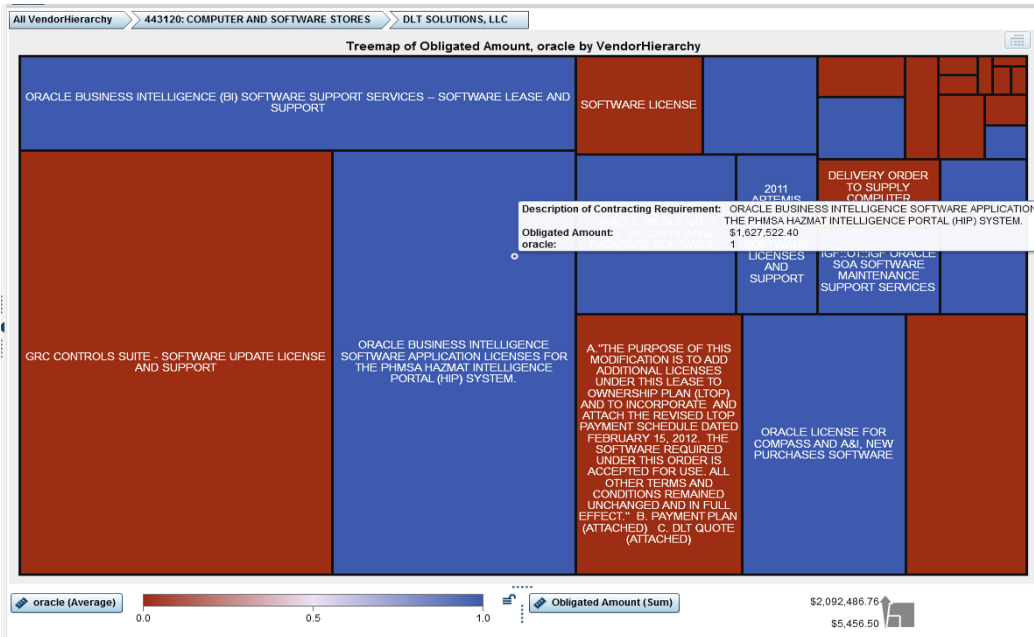


Figure 9: Tile Chart Highlighting Oracle Spending Where DLT Solutions LLC Is the Vendor

SAS Visual Statistics, an add-on module to SAS Visual Analytics, also produces insight into the context of large spending initiatives from the DOT. Figure 10 depicts just one feature of SAS Visual Statistics, detailed decision trees against continuous variables. By selecting a target variable such as spending amount per line item, we can gain insight into how various structured data tend to impact spending, including the newly created structured fields indicative of spending on a particular software initiative or with a particular vendor. The decision tree can be depicted in a variety of ways, and for this data set, a number of buckets appear indicative of higher spending amounts. For example, given definitive contracts associated with Federal Aviation Administration and the Federal Transit Administration, for a particular set of spending codes including Computer Systems Design Services, there tends to be a focused amount of larger transactions. Such methods enable analysts to further explore these combinations of variables in an effort to understand and potentially consolidate extraneous spending.

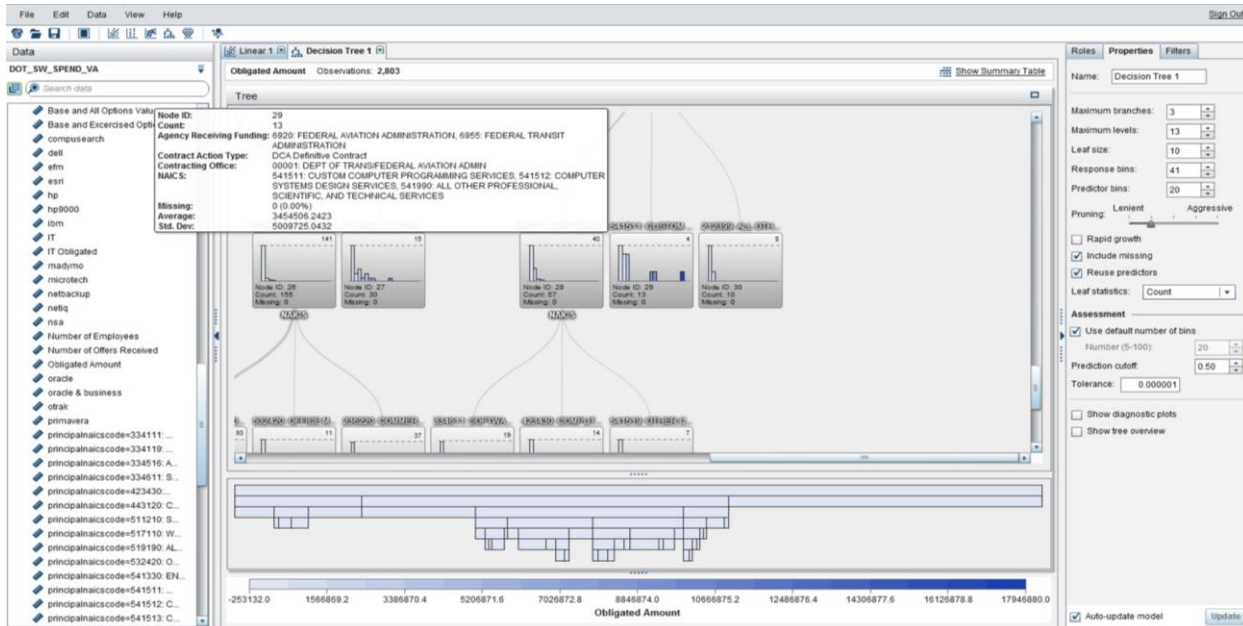


Figure 10: Decision Tree by Spending Amount in SAS Visual Statistics

INTERACTIVE REPORT GENERATION AND USE

Finally, SAS Visual Analytics enables the generation of interactive reports against the available structured and unstructured data. The report benefits greatly from interactive filtering. The images below depict how we enabled selection of particular software solution vendors that emerged from the text analytics exercise, including SAS, IBM, and Oracle. Then, analysts are able to explore how each of these software solutions/vendors are deployed across the DOT. In Figure 11 below, we depict how IBM software products span the Federal Aviation Administration, highlighting \$1M+ spending on the SOAR project. In Figure 12, we depict how Oracle software associated with a particular delivery order and the NAICS code associated with Computer and Software Stores correlates with a large investment in the Iron Bow technology vendor in December 2011.

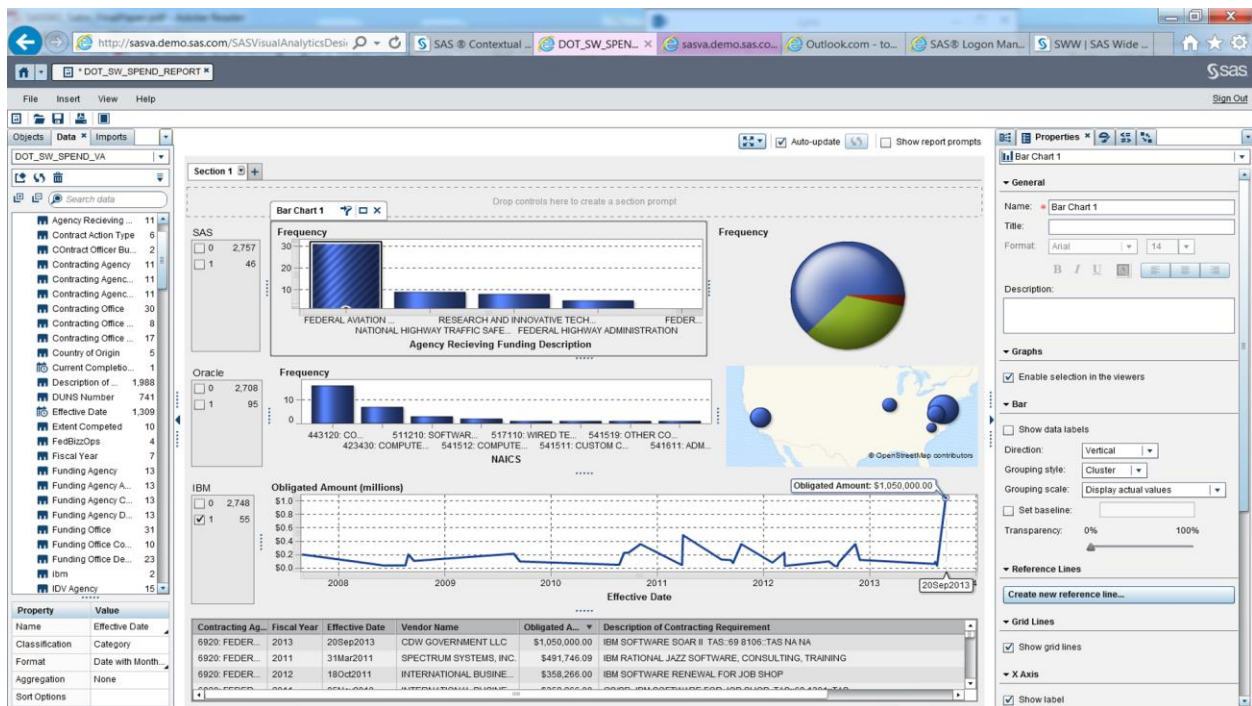


Figure 11: Interactive Reporting Using SAS Visual Analytics Reporting against IBM Software Spending

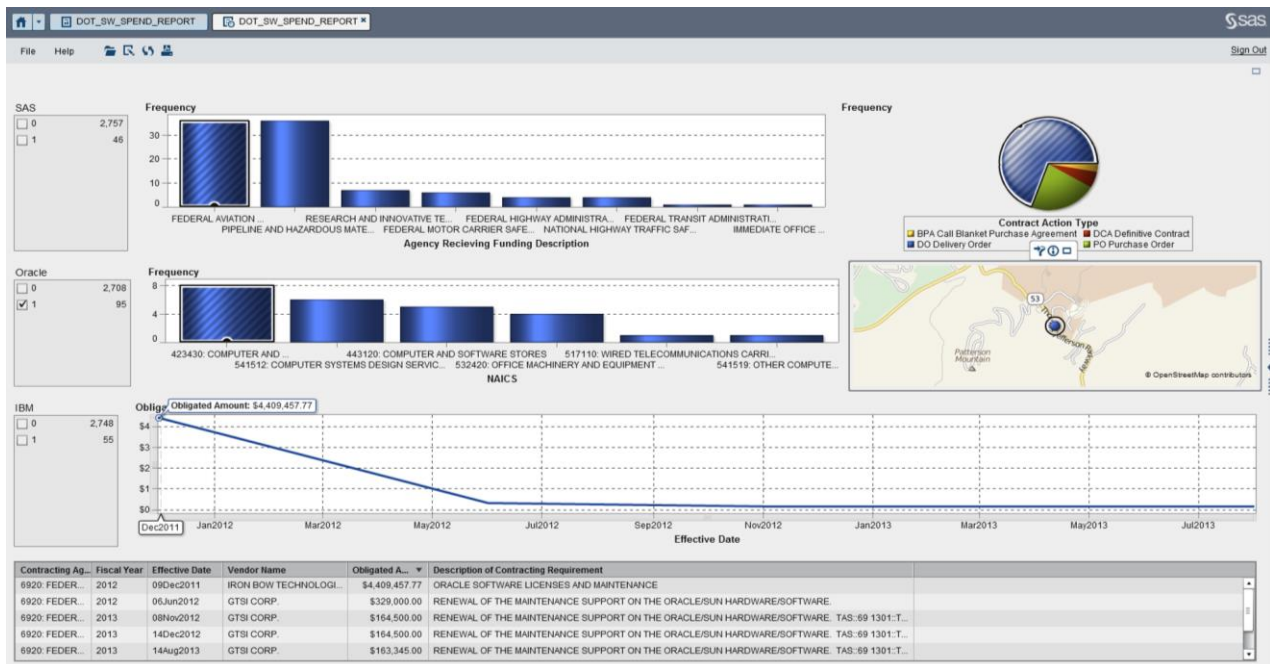


Figure 12: Interactive Reporting Using SAS Visual Analytics Reporting against Oracle Software Spending

DISCUSSION AND CONCLUSION

In this paper, we have expanded upon a previous global forum submission² to further demonstrate a repeatable framework for research analytics, this time focused on applying the framework to government spending. We obtained spending data associated with a particular federal government agency, the Department of Transportation, from a public source, and extracted data associated specifically with software spending using SAS code to flag such data. Within SAS Contextual Analysis, we were able to use a categorical variable to generate text rules which differentiated particular software applications, vendors, and projects by spending code. We prepared the data and surfaced the new text analytics binary categorical variables alongside the original structured and unstructured data in SAS Visual Analytics and SAS Visual Statistics for ad hoc exploration, modeling, and interactive reporting.

In the process, we showcased how the text analytics results can be used to identify multiple vendors associated with software solutions and packages. We were able to use models and interactive reporting to highlight key characteristics of large projects and initiatives. We also saw how exploring the data in an ad hoc manner highlighted potential data quality issues, such as the wrong spending code ascribed to SAS, as well as multiple names for seemingly the same vendor. Finally, the interactive reporting capabilities and dynamic filtering provide a way to quickly ascertain extensive software spending, and tie this spending back to software applications, vendors, and certain projects identified from the unstructured data.

The capabilities highlighted in the paper just scratch the surface of what is possible using text analytics and subsequent visual analytics and exploration against this data. Here are suggested follow-up projects for those with the initiative:

- Apply to different/multiple/all government agencies. This will require a scalable high-performance infrastructure.
- Further explore the interplay between classification based and exploratory based text analytics. For instance, we could start with a categorical definition of Oracle spending based on what is provided to us by an organization's OCIO. This may include a number of products not designated by the term "Oracle". Subsequently, we can perform exploratory text analytics through Contextual Analysis against spending scored with this categorical definition, looking for connections with other vendors, software solutions, and agencies.

- Apply topic identification and extraction. This can be done at a wide scale, as up to 999 topics can be extracted using SAS Contextual Analysis. In previous demonstrations, we have uncovered a wealth of information in the topics extracted. For example, one topic identified against this data represented one of the largest projects by spending at the DOT. Expectation maximization algorithms could highlight topics around large spending.
- More visualization options are available given a single categorical variable, which could be generated from the text analytics. In this exercise, we produced 200+ binary variables based on the text analytics results. Most of these do not overlap, though some do. It would be worthwhile to explore compressing these 200+ variables into one categorical variable, combining identified software applications and vendors into one entry where necessary, or finding some method to highlight one versus another. Also, determining ways to use the 200+ binary variables in conjunction with this new categorical variable would be worth exploring.
- Further exploring predictive capabilities against this data would be worthwhile. The rule builder capability of SAS® Text Miner could be used to generate a number of rules similar to what is generated from SAS Contextual Analysis, but in a single categorical variable with precision/recall parameters. These could be used to guide visualizations by topics, for example, interactive reports by category, which highlight particular software applications and rule combinations of software, projects, and technology.

Text analytics continues to evolve at a rapid rate, and the ability to incorporate these capabilities into initiatives involving unwieldy amounts of text is paramount in many domains. For more insight into the use cases within government where text analytics can be applied, please investigate the paper “Text Analytics in Government” highlighted in the references below.¹

REFERENCES

1. Sabo, Tom. 2014. SAS Institute white paper. “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data.” Available http://www.sas.com/en_us/whitepapers/text-analytics-in-government-106931.html.
2. Sabo, Tom. 2014. “Uncovering Trends in Research Using Text Analytics with Examples from Nanotechnology and Aerospace Engineering.” Cary, NC: SAS Institute Inc. *Proceedings of the SAS® Global Forum 2014 Conference*. Available <http://support.sas.com/resources/papers/proceedings14/SAS061-2014.pdf>.
3. USAspending.gov. Available <http://www.usaspending.gov>.
4. volpe.dot.gov. Available <http://www.volpe.dot.gov/about-us>

ACKNOWLEDGMENTS

Thanks go out to Beth Ebersole and Peter Paschalides for both providing the data utilized for this investigation, and for doing significant work with analytics and visualizations against the structured data. Manual investigations of certain software solutions from the unstructured contract description field led to the effort described in this paper to automate the process. Also thanks to Manuel Figallo, Emily McRae and Peter Paschalides for reviewing this paper and, in the process, suggesting avenues for further investigation.

CONTACT INFORMATION

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

Name: Tom Sabo, Principal Solutions Architect
 Organization: SAS Federal LLC
 Address: 1530 Wilson Blvd.
 City, State ZIP: Arlington, VA 22209
 Work Phone: +1 (571) 227-7000 x51717
 Email: tom.sabo@sas.com

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.

Other brand and product names are trademarks of their respective companies.