A Stand-Alone Methodology for Data Exploration

in Support of Data Mining and Analytics

A Thesis

Presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Industrial Engineering

By

Michael W. Gage

May, 2013

COMMITTEE MEMBERSHIP

TITLE:	A Stand-Alone Methodology for Data Exploration in Support of Data Mining and Analytics
AUTHOR:	Michael W. Gage
DATE SUBMITTED:	May, 2013
COMMITTEE CHAIR:	Dr. Reza Pouraghabagher, Professor Industrial & Manufacturing Engineering Department, California Polytechnic State University, San Luis Obispo
COMMITTEE MEMBER:	Dr. Jose Macedo, Professor Industrial & Manufacturing Engineering Department, California Polytechnic State University, San Luis Obispo
COMMITTEE MEMBER:	Dr. Roya Javadpour, Professor Industrial & Manufacturing Engineering Department, California Polytechnic State University, San Luis Obispo

ABSTRACT

A Stand-Alone Methodology for Data Exploration

in Support of Data Mining and Analytics

Michael W. Gage

With the emergence of Big Data, data high in volume, variety, and velocity, new analysis techniques need to be developed to effectively use the data that is being collected. Knowledge discovery from databases is a larger methodology encompassing a process for gathering knowledge from that data. Analytics pair the knowledge with decision making to improve overall outcomes. Organizations have conclusive evidence that analytics provide competitive advantages and improve overall performance. This paper proposes a stand-alone methodology for data exploration. Data exploration is one part of the data mining process, used in knowledge discovery from databases and analytics. The goal of the methodology is to reduce the amount of time to gain meaningful information about a previously unanalyzed data set using tabular summaries and visualizations. The reduced time will enable faster implementation of analytics in an organization. Two case studies using a prototype implementation are presented showing the benefits of the methodology.

Keywords: data mining, analytics, visualizations, data exploration

ACKNOWLEDGEMENTS

I would like to thank the company which provided the anonymous data for the second of two case studies presented in this thesis. I would also like to thank Bill Gage for his tireless support and discussion about implementation best practices, Fairy Gage for her advice and perspective about my work, and Katherine Gage for her encouragement throughout the process of brainstorming, development, and writing.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (continued)

Page

TABLE OF CONTENTS (continued)

Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES (continued)

Figure Page

CHAPTER 1

Introduction

The purpose of this thesis is to design and promote a systematic stand-alone methodology for analysts to use in order to increase the effectiveness and speed of their data analysis. An implementation of the methodology will also be provided with two case studies of its applicability. This methodology seeks to fill an area where few improvements have been made, in spite of tremendous progress in other areas related to information management, data mining, and analytics.

Big data is the type of data that is consistently being collected and stored in large quantities across a variety of formats. As a result, Big Data has presented challenges to every aspect of information management. Many hardware improvements and new software tools have been developed to facilitate faster data storage and access times, an increase in storage space, an increase in the variety of data stored within a single database, and improved analysis performance. Databases used for collecting data are often cloned into data warehouses to separate the steps of collection and retrieval into two different systems. This allows the data to be stored in the ideal format for both collection and retrieval, avoiding the performance problems of a single system handling both types of transactions.

The process of knowledge discovery in databases (KDD) has been the main methodology developed to address gaining insight into all of the collected data. The first few steps of the knowledge discovery methodology are often performed by the data management systems to prepare and transform the data into the appropriate format for analysis. The most complex step of KDD is data mining. Data mining often involves regression, cluster analysis, and other specific methods which seek to aggregate and summarize the large volumes of data into meaningful information.

The most common way to share or summarize the results from data mining is through data visualization. Despite the increase in performance of both software and hardware to match the challenges of big data, the relative ability of people to comprehend volumes of information has not increased significantly. Through carefully designed intricate graphics, a data set of millions or billions of records can be summarized and a picture of the information within the data can be painted. Visualizations are computationally expensive to create, but significant progress has allowed the possibility of multi-dimensional interactive graphics.

Analytics is the business application of the knowledge and information contained within large data sets. Analytics brings together the results of KDD and communicates those results as a specific action to be taken. Analytics are often used in corporations to drive performance improvements; first collecting data about processes, then analyzing the data to provide information using KDD, and then taking the information and applying it to make changes to the original processes to improve the results. Visual analytics are the combination of analytics and data visualization, where the results of KDD are summarized through a frequently updated visualization of the process performance. The visualizations display key performance indicators for an organization with the most current data.

Many large companies such as Oracle, IBM, EMC, TeraData, and Tableau have created enterprise tools to perform KDD, analytics, and create visualizations. These tools are part of suites offered by the respective companies which must be integrated into the information technology architecture of any organization to be used effectively. The users must have a moderate understanding of a wide variety of computing topics, as well as knowledge about advanced statistical analysis and understanding of the data which they seek to analyze.

These advances and tools have all sought to improve the amount of information which can be calculated and resulting knowledge gained from the collected data by improving hardware and software systems. However, analysts, managers, and other interested parties may not have access to the systems required to support these tools and the background understanding of computing topics and advanced statistical analysis. For these individuals, who are numerous in organizations around the world, the best tool in their suite of understanding continues to be spreadsheet software with stagnant data and graphs generated one-by-one through a menu system.

The methodology proposed by this paper seeks to fill the gap between the commonly used spreadsheet software and more advanced enterprise tools. The methodology presented defines a standardized process for crossing the gap between the tools, separate from any other tools, as a stand-alone solution. Data can be obtained from any source in a specific format and is automatically analyzed for basic information which may assist with data modeling, cleanliness, and initial understanding of previously unanalyzed data. The user is also asked to provide some information which is commonly provided in graph wizards. The information is collected about every column of data so further input will not be required. After the initial basic information is provided, further generic analysis of the entire data set is performed to obtain and provide basic statistics.

The data is grouped using values within columns identified as independent variables to provide more in-depth information than simple spreadsheet tabulations. The results are provided in both a graphical and tabular format so the user can see the information and the supporting summary.

The goals of this methodology are to assist any user on a desktop or laptop computer in the analysis of data about which he or she has little or no information. The time the user must spend to gain meaningful information about the data should be reduced significantly. A user looking for specific information about a data set may see a trend in a visualization or data summary outside the scope of the original investigation. These insights allow the user to explore the data leading to further insights.

CHAPTER 2

Background & Literature Review

Big Data

The discussion of any data analysis must begin with the type of data to be analyzed. The term Big Data has recently been coined to describe the vast quantities of data that are collected and stored on a daily basis around the world for various purposes. Some characteristics which help to understand Big Data include that it is an ongoing continuous collection over time and the data collected from a wide variety of sources can be meshed together in a rapidly increasing number of ways to gain further information.[23][22] Three characteristics which are often used to describe Big Data are referred to as the three Vs: volume, variety, and velocity.[14][22]

The volume of information is hard to grasp or understand through typical analogies of size. Relevant descriptive names for large volumes of data include an exabyte, equivalent to 2^{60} bytes, a petabyte, equivalent to 2^{50} bytes, a terabyte, equivalent to 2^{40} bytes or one trillion bytes, and a gigabyte, equivalent to 2^{30} bytes or roughly one billion bytes. There are estimates the amount of data created each day around the world is in excess of 22 exabytes.[35] An analogy which helps convey the magnitude of the volume is if one gallon of water in the Atlantic Ocean represented one byte of data, the data generated by the world in 2010 would barely fit in the Atlantic Ocean.[14] These figures immediately beg the question of where all of the information is being stored in such high volumes. Many companies are creating data centers which can store data in excess of 50 petabytes or 50 thousand terabytes of data.[35] In addition to the data centers, there were an estimated one billion personal computers in use around the world in 2008.[34] This amounts to a very large storage space, though it is widely spread out across data centers and the personal computers.

The variety of information, while not rapidly increasing, already comes from many diverse sources. The biggest differentiating factor is between structured and unstructured data. Structured data contains numbers, dates, or other information which are limited in size and consistent in format across various sources. Unstructured data is highly variable in size, format, and source, and includes pictures, videos, and text. Structured data, which is much easier to process, is the tip of the iceberg of all possible data; it is estimated to represent only 10% of the total data generated, with the remaining 90% being unstructured.[35] Some common consumer products which create large volumes of data include Facebook, Twitter, MySpace, Google+, and Pinterest.[22][35] The data includes posts, GPS tags, pictures, videos, and relationships between individuals. Corporations collect information such as banking transactions, employee performance, and salary information.[22] Public census data, including a wide variety of structured content, is collected around the world.[27] Images from scientific research in genomics, biochemistry, neurology, astronomy, and physics are some of the largest sources of data in existence.[23]

The velocity of information is equally challenging to grasp in comparison to the volume and variety of data. The one billion personal computers in use as of 2008 are expected to double by 2014.[34] Individual devices, including personal computers and cell phones, represent one of the biggest challenges to the speed at which information is collected. Users can be logged into services like Facebook from multiple devices while

Facebook is recording information about where they are logged in from. At the same time Facebook also provides information from hundreds, if not thousands, of other users to each user on each device.

With all of this data being collected, it only makes sense that it be put to use in some way rather than only collected and then stored indefinitely. The most challenging aspects of Big Data are that the data must be "ingested, processed, aggregated, filtered, organized, and fed back in a meaningful way for businesses to get some value out of it."[14] Many contend that "transaction processing and data storage are largely solved problems" and the true problems Big Data presents are "primarily those of analysis."[23]

Knowledge Discovery from Databases

The automated analysis of Big Data has been developed into a framework or process known as knowledge discovery in databases (KDD).[7] It is considered to be an iterative process with the following six stages:

1. Develop an understanding of the proposed application
2. Create a target data set
3. Remove or correct corrupted data
4. Apply data-reduction algorithms
5. Apply a data-mining algorithm
6. Interpret the mined patterns

The second and third steps are often grouped together and referred to as pre-processing.[29]

The first step of KDD, developing an understanding of the proposed application, is crucial to realize why the KDD process is performed. Some information which may be contained within the data is of interest to a user or users. This is usually very specific, based on the perspective of the user, and helps to limit the problem to an analysis of greater depth, rather than one of greater breadth.

The second step of KDD, creating a target data set, is often performed in conjunction with data collection and is in the form of data warehousing. Data warehousing is the cloning of data from a database which is collecting or recording transactional data into a repository, alongside adjacent data from other databases which may be relevant.[14] Data warehousing separates the challenging step of analysis from the need to process incoming data, allowing the flow of incoming or changing data to remain uninterrupted. A data warehouse, by nature, at least doubles the amount of storage space required.[14] In some cases, for the purpose of analysis, relational databases are denormalized to improve the efficiency of analysis algorithms, resulting in cross products of relational tables and a much larger footprint than the original database.[23] The less efficient storage compounds the problem presented by Big Data in terms of volume. As mentioned, the hardware to store data continues to increase dramatically in capacity to accommodate this data.

The third step of KDD, correcting or removing corrupted data, is fundamental to gaining accurate results. The correction of corrupted data should arguably feed back into both the original databases and data warehouses to correct the source or sources of the corrupted data.

The fourth step, reducing the data, is also commonly referred to as transforming the data. This is especially important when dealing with extremely large data sets to effectively reduce or summarize the data prior to the automated analysis. A data warehouse often provides a comprehensive list of attributes about any given item of interest, usually in a universal or denormalized table. For the purposes of the data mining algorithm, any excess columns should be excluded to provide improved performance and avoid the possibility of noise within the results.[29]

The fifth step of the KDD process, data mining, is often referred to rather than KDD itself. It is the engine behind the results of the processes. Further discussion of data mining is needed and presented after the remainder of the KDD process.

The sixth and final step in the KDD process usually involves further filtering and summarization of the outputs. Many results from KDD and data mining models do not apply to any application and are meaningless or provide no new information. Other times the results may not apply to the particular application, but can be relevant to other applications and may provide information that leads to knowledge about a particular subject.

Data Mining

As the engine of the KDD process, data mining receives an increased focus. With this focus, data mining has the following six main functions to gain information from data.

1. Classification – finding models that analyze and classify a data set into several predefined classes
2. Regression – mapping data to a prediction variable
3. Clustering – identifying a finite set of categories (not predetermined) to describe the data

4. Dependency Modeling or Association Rule Learning – finding a model which describes significant dependencies between variables
5. Deviation Detection or Anomaly Detection – discovering the most significant changes in the data
6. Summarization – finding a compact description for a subset of data

A wide variety of potential methods or algorithms perform these functions, including decision trees, neural nets, database segmentation, market-basket analysis, and strict deviation detection.[7]

Data mining is the most emphasized step of the KDD process, because choosing a data mining method determines what type of results will be delivered. The choice of a method and the process is separated into six additional steps shown in Figure 1, defined as the Cross Industry Standard Process (CRISP) for data mining.[29]

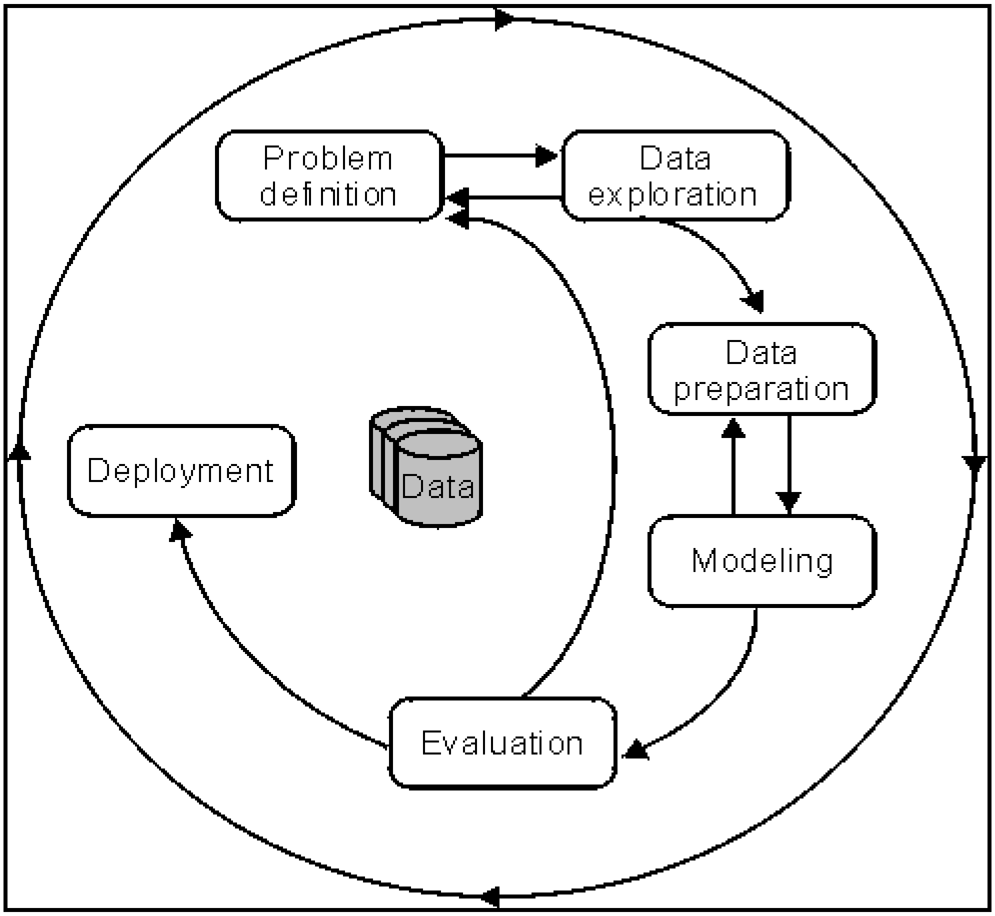


Fig. 1. CRISP-DM Process

The six steps are defined as follows.

1. Problem Definition – understanding the business or real world problem which poses a question to be answered
2. Data Exploration – understanding the data, providing a good description of the data, including statistics and identifying quality problems in the data
3. Data Preparation – collecting, cleaning, and formatting data, including the derivation of additional attributes based on information contained within other data attributes

4. Modeling – applying various mining functions
5. Evaluation – verifying the model addresses the problem and takes into consideration all possible relevant issues
6. Deployment – feeding the results back into database systems or other applications

The data exploration stage of this process has an understated role in the steps of data mining and in the process of KDD. The task of exploration is left to “traditional data analysis tools”, despite the acknowledgement that entirely different methods are applied throughout the rest of the process. Spreadsheets cannot handle the volume of data contained within databases; therefore the data is segmented accordingly and manually analyzed. Additionally, domain experts may not be familiar with statistical analysis tools, requiring a statistician to become familiar with the domain first before exploring the data set. This challenge presents an opportunity for improvement in this stage of data mining. The methodology proposed by the author seeks to take advantage of this opportunity.

KDD and data mining were originally designed for scientific research applications to gain information but not to guide the actions of an organization. Extensions to the KDD process are needed to define how executives and managers can use the information to make better informed decisions.

Analytics

One well developed extension to the KDD process is that of analytics, often referred to as advanced analytics when used in conjunction with KDD. Although the KDD process identifies patterns and relies on the domain experts to identify whether those patterns are meaningful, there is no further step in the process to use the knowledge acquired in the process. Advanced Analytics is defined by Forrester Research Inc. as "any solution that supports the identification of meaningful patterns and correlations among variables in complex, structured and unstructured, historical, and potential future data sets for the purposes of predicting future events and assessing the attractiveness of various courses of action."[26][25] The key differentiator between KDD and analytics is the assessment of various courses of action. Analytics are often employed by businesses, organizations, and governments to determine meaningful patterns and then adjust their courses of action based on the discovered patterns. The results of KDD are "used to direct, optimize, and automate their decision making."[5] The format and means for communicating the results varies and presents an additional opportunity for making the information more meaningful.

Data Visualization and Visual Analytics

One of the most effective means of communicating the results from KDD is through visualizations of the information. Though it is often overlooked, "the best pattern detector we have at our disposal – the human visual system" can be put to use to gain information from data.[3] "Visual representations and interaction techniques that exploit the human eye's broad bandwidth pathway into the mind to let users see, explore, and understand large amounts of information simultaneously" makes data visualization an ideal means of communicating when it comes to presentation and analysis of Big Data.[11] The cliché 'a picture is worth one thousand words' could be rewritten in the case of data visualization. Visualizations are worth the enormous amount of background data used to create them.

Visual analytics is defined “as the science of analytical reasoning facilitated by interactive visual interfaces.”[11] Analytics can be paired with intuitive graphics and interfaces to increase the amount of information and the speed at which information is understood by an individual. The goal of analytics is to provide key information to improve performance across levels of a business or organization. This goal is greatly facilitated by visual aids. Visual analytics can provide “techniques to support production, presentation, and dissemination of analytical results to communicate information in the appropriate context to a variety of audiences.”[12] Detail oriented analysts are focused on specific information contained within data. High level decision makers are looking for concrete information to run their organizations. Allowing information gained from a complex data analysis process such as KDD to be presented in simple graphics brings together the analysts and decision makers.

Although visualizations seem to be an ideal solution for the presentation of results from KDD as a part of analytics, there are some drawbacks to consider. The first and foremost of these is the additional computational challenges complex graphics present. The computational challenge compounds the already difficult problem of analyzing Big Data. Digital and print displays of information are limited in terms of resolution and cannot adequately present even significantly summarized models of some large data sets. The loss in performance of algorithms providing information is considerable when colors, textures, and multi-dimensional shapes are used to render information and patterns uncovered in a data set. Simply the task of providing multimedia content based on such large data sets over wired and wireless computer networks compounds the problems of Big Data itself. The interface which conveys a visual and interacts with a user is also a significant challenge. When an individual is navigating using a mouse or touch screen interface through a three dimensional data set, there are limitations to what the user can do.[10] Any three dimensional interaction uses a combination of mouse and keyboard actions or buttons to simulate the ability to change perspective.

The computational challenges of creating, transferring, and interacting with visualizations aside, visualizations occasionally present the information in qualitative ways, which can lead to multiple interpretations. Visualizations can be even more misleading than quantitative statistics used in news stories. For interactive visualizations, the degree to which a user interacts can also change the amount or types of information gained. A ‘lazy' user may pull up a graphic and take it at face value rather than interacting with it. Other users may find vital information and manipulate the visual in such a way they are unable to determine how to return to the information they originally saw. Visuals are also very reliant on the context of the data they represent. A simple case of context relevance is in the proper use of a different chart types. Pie charts best demonstrate proportions of a population while time series charts show trends over time. If the context and visual do not match, like a pie chart of time series data or a time series chart of proportions, the information loses value immediately.[28]

An important note is that "visual techniques provide a first line of attack that can suggest what kind of trends and patterns may lie within the numbers, and may thus help guide the focus of more detailed analysis."[7] This type of approach is known as an exploratory analysis, similar to the data exploration sub-step in the data mining step of the KDD process. Visualization tools are not often considered at this stage due to the proprietary nature of commercial visualization tool algorithms and the lack of software development tools enabling the creation of visuals. There is a very common misconception that one must “use complex visualizations to solve complex problems.”[9] Complex visualizations can be valuable to solve a complex problem, however simpler visualizations prove to be even more useful. Simple questions can be answered very quickly in very easy to understand visualizations.[9] The methodology presented by the author seeks to use easy to understand visualizations to answer questions and problems with a variety of difficulty.

Existing Tools and Applications

Tremendous progress has been made in providing applications and tools to store and manage Big Data, data warehousing, data mining, KDD, analytics, and data visualizations. Companies including Microsoft, Oracle, IBM, Dell, and Informatica have developed specific suites of technology to handle the storage and retrieval of Big Data.[14] FICO, IBM, SPSS (IBM), Angoss, KXEN, Oracle, Portrait Software, Rapid-I, SAS, Microsoft, Teradata, TIBCO Software, Business Objects, DBMiner, MicroStrategy, and WebTrends have all developed software tools for data mining.[5][26] Minitab, SPSS (IBM), SAS, StatSoft, and TIBCO Software have developed software designed for statistical analysis and, in some cases, connectivity to databases.[26] Tableau, DataDog, Qliktech, and Edgespring have all created and distributed visualization tools to organizations.[14] Tableau has begun to integrate further analysis tools into its software to allow its application to be integrated with existing database systems similar to products provided by SAS.[9]

All of the above tools are licensed software, marketed and sold to organizations on a daily basis. Pricing information is only available through a direct quote from a consulting sales team and varies depending on the application. Pricing has been rumored to be $20,000 for a single user with prices declining for group licensing.[36] Often the reason for selection of a particular tool or set of tools is related to integration of the tool or tools into existing information management architecture, or the desire to avoid additional integration effort and expense. Unfortunately, many organizations are often forced into hiring experts across computing disciplines to configure the environment and analysts with detailed knowledge of the information management architecture. There are some solutions which operate without formal information management architecture, such as Tableau. However, the capabilities of those software solutions are limited to a great extent without a cluster of servers and the networking infrastructure to improve performance and support the operations.[23][9]

Two tools which are open source and freely available for use are RapidMiner from Rapid-I as a data mining tool and R from the free software foundation as a statistical software and visualization tool.[26] RapidMiner is a complete package, but as free software does not have any support other than available user forums. RapidAnalytics is the paid version of the software from Rapid-I and includes additional features and support.[30] R is a very powerful tool which is completely open source and ever expanding in terms of repositories and plug-ins. However, the extensibility and programming focus of R creates a very steep learning curve for users who are trying to use R for analysis. Often entire customized packages for R are developed for a single specific application. These packages are available to other users but are rarely applicable or useful for more than the originally intended purpose.[8]

The performance characteristics of many of these tools have been investigated for both qualitative and quantitative benefits. Quantitative benefits include memory usage, percentage of memory and paging, page file usage, disk space, and disk access times. Qualitative benefits include hardware support, connectivity, algorithmic variety, standardized methodology, user interface, visualization, data cleansing, and binning capabilities. Results from these studies show a very competitive market between different vendors, with slight advantages and disadvantages depending on the organization, the hardware available, and the type of tasks being performed.[1][19] A major topic of discussion and a choice by some tools, including R, is to perform tasks in memory only. This presents a significant limitation in terms of handling Big Data and creating visualizations. Old architecture design issues prevent the full utilization of modern hardware improvements in software tools. The transition from 32-bit to 64-bit is the most significant architecture design limitation preventing full memory addressing and efficient memory utilization.[23][19]

Reasons for Adoption of Analytics

Analytics have proven to be a tremendous resource to companies and organizations. Typically these decision making systems have been used in "financial forecasting, budgeting, and supply chain management."[24] From 2010 to 2011, there was an increase of 21% in the number of companies who report gaining a significant competitive advantage from using analytics [24].

Organizations' adoptions of analytics have been classified into three main categories: aspirational, experienced, and transformed. They are classified based on their information management policies, analytical skills and proficiencies with existing tools, and the extent to which their culture is data-oriented. These classifications were selected to view how well organizations manage, understand, and act on data. Aspirational organizations have limited information management, limited analytical skills and proficiencies with existing tools, and a culture that is not data oriented. Experienced organizations fall into two groups: collaborative and specialized. Collaborative organizations are focused on information management across the enterprise with a cultural emphasis on using data to make decisions, but less developed analytical skills and tool proficiency. Specialized organizations have well developed analytical skills, but information management is isolated and the culture is not focused on data. Transformed organizations are proficient, if not advanced, in analytical skills and tools and have information management and a cultural emphasis on data-driven decision making across the enterprise.[17][24]

The longitudinal study of these groups has shown that a large gap is forming between the two groups using analytics regularly, experienced and transformed organizations, and the group which does not use analytics, aspirational organizations. Aspirational organizations are continuing to decline in their belief that analytics are providing a competitive advantage, while both experienced and transformational organizations have continued growth in the belief that analytics is providing them a direct competitive advantage. In all groups, the organizations using analytics to gain a competitive advantage were more than twice as likely to outperform their peers in industry.[17]

Many of these shifts and the applicability for analytics has resulted in economic and market volatility and positioning. With large volumes of data providing the backbone for decision making, decisions can be made faster to avoid potential losses and to seize opportunities which would otherwise go unnoticed. Day-to-day operations in transformed organizations are almost entirely automated with management and analyst oversight, allowing the focus of the business to shift to long term priorities.

Analytics provide improved future planning and operational risk assessment, a cascade effect from the shift away from day-to-day operational focus. With less effort spent on daily planning, analytics can be used to identify possible future pitfalls and opportunities based on the huge volumes of data being stored each day. The additional information about risk is used to change strategy to be more risk-averse or to include contingency plans for more potential risks.

Analytics also improve focus on the customer. Transformed organizations are collecting more data about their customers' habits, needs, wants, and feedback. In turn, the data collected is analyzed and the resulting information is put to use to make changes to engage the customer more effectively. Determining patterns in the activities of various customers can identify cues to improve an organization's ability to retain existing customers and gain new customers.

For any organization which is attempting to improve its stance towards analytics, there are steps to be taken and pitfalls which present risks. The companies surveyed identified organizational challenges to be nearly twice as difficult to overcome or resolve as issues surrounding technology for the purposes of analytics.[17] Based on this, information management and developing a culture of focused on data driven decision making should be the focus of any organization trying to improve. Without data being collected and information from that data being shared, no analytical tools or skills can be an asset to an organization. With solid information management practices, analytics skills can be taught to employees on a gradual basis while acquiring external support. Analytical tools can always be purchased or investigated, with a focus on tools that work well with the existing information management and infrastructure.[17] Analytical tools can immediately be leveraged once purchased in an organization with good information management practices. With the understanding of the existing methodologies, processes, and tools, potential applications to specific fields of study are needed to assess the potential for new developments.

Areas of Application

There are a wide variety of areas which face the challenges presented by Big Data, and many of those areas correctly note opportunities for improvement when analyzing and putting Big Data to use. Each area presents its own unique shift from the existing forms of data collection and analysis. The following are a small sample of the areas which could benefit from analytics.

Human Geography

One area which has recently felt the impact of the connected nature of society is human geography. In the past, geographers spent time obtaining survey data and struggled to obtain information about the movements of people around the globe and the shift of populations. Now, these geographers face a seemingly infinite flow of data from smartphones, emails, social media posts, photographs, and videos from around the world, which all have geospatial information attached. One example is the information that is now recorded by swipe card machines in London. Seven million passengers board public transportation using this system each day. This comes out to roughly 2.5 billion trips that are taken each year, which is steadily increasing.[4]

Astronomy

The field of astronomy has felt the impact of Big Data arguably before other fields. In the early eighteenth-century, roughly 3000 stars had been cataloged. By the late nineteenth century, roughly 300,000 stars had been cataloged. More recently, since the 1990's, the Sloan Digital Sky Survey has been taking photographs of the sky, generating 200 gigabytes of data each night amounting to roughly 50 terabytes of data by 2012. Not only has the survey generated imaging, but 3000 papers on various topics have been cited 160,000 times. The next large project, the Large Synoptic Survey Telescope, is expected to take images amounting to 5-10 terabytes of data each night, storing that information in a 60 petabyte database. Astronomy is a more mature field in its acknowledgement of the need for computers to process data automatically to identify planets and stars. Despite the extensive use of data mining and other tools in the field, the 20 billion rows of data in the Large Synoptic Survey Telescope database outpaces any advances which have been made in improving the computationally expensive analysis.[2]

Genomics and Medicine

Other fields which have felt the influx of data are genomics and medicine. As of 2012, the amount of data that can be generated sequencing the human genome in one day exceeds the data the entire Human Genome Project created from 1990-2003.[21] Geneticists have identified statistical anomalies in such large data sets, a problem which will affect other fields as well. Those anomalies cannot be confused as real phenomena.

Neurologists who look at brain imaging have a great deal of data and are struggling to unlock the full potential of the information residing within the images. The three dimensional structure of the data about each brain collected over time presents a challenge to analysts of how to compare the data that is being collected to discern patterns. Another longitudinal project is working to collect and analyze the medical records of 100 million patients to monitor drug and medical device safety. Information about blood pressure, lung capacity, antibody concentration, blood work, imaging, and other data is collected at every visit to any type of doctor by those patients. A challenging aspect of this work is the effect of individuals who do not regularly visit doctors or who drop out of the study. Whatever model is used in data mining will have to account for these missing values, while still taking into account the remainder of patient data.[31]

Politics

Politics around environmental data are also being impacted by Big Data. A more and more common argument between politicians is how much funding should be spent on collecting data about the environment. When budgets shrink, the funding supporting more data gathering activities is one of the first items to be cut. This is not the same as the budget for astronomers who have as much data as they require. Outside of the problem of data gathering is funding and support for the analysis of environmental data. Within that issue is the potential problem of political influence and misuse of the results from analysis. People with particular political views have consistently misused statistics to sway public opinion and information. The decision to collect data has to be supported by a supposed problem to receive funding. Compared with other fields, this stance is backwards and does not make sense; it is the opposite of a data driven culture. Businesses and individuals are also fighting to prevent public and political access to data, citing proprietary operational data cannot become public. Intellectual property rights are at the forefront of this debate, and politicians have gone so far as to discredit scientists who published results from studies without the original data. The political arena is the most ethically challenging area for collecting data and implementing practices to analyze Big Data.[18]

Transportation Safety

The Center for the Management of Information for Safe and Sustainable Transportation (CMISST) has recently expanded the research it is performing on different traffic data. Crash, emergency services, and hospital data is being used to determine the effectiveness of highway-safety in conjunction with emergency services to prevent fatalities. The goal is to expand this to include serious injuries. The field performance of safety technology can be evaluated by linking crash data to vehicle safety features. The effects of roadway characteristics on crash occurrence and outcome can be viewed in a much broader scale using data collected around the country on a daily basis. The most interesting potential use of the data is linking crash data and law enforcement patrol activities to study the effectiveness of the enforcement techniques.[33]

History

Records of history provide an opportunity for a newer area of Big Data. Historians can create a profile of text documents written over time to demonstrate how well events are remembered and how quickly they are forgotten over time. This provides insights into differences between the perception of what was important during a time period and what stood out to historians as important from the time period. A creative visualization of ship logs is one of many unique ways historians are revisiting the variety of data they have to observe patterns and share the information with others. The visualization shows the exploration of the world and transportation of people around the world. Researchers at Google have taken to looking at current news stories to determine the effects of new technologies, such as social media, on the length of time famous people stay in the news. Interestingly enough, the length of time a famous person remains in the news has changed very little over the course of history, despite the increasingly wider array of access to information. Copies of letters, diaries, newspapers, and public records from the Civil War can be analyzed and put together to learn entirely new information about the time period. A more recent project is taking data about railroads in the form of cartoons, poetry, maps, timetables, and abandoned track lines to understand more about how the railroad was built across the country. Additional data from annual reports, censuses, newspapers, and other scholarly writing was joined alongside the initial data to obtain a very accurate understanding of the railroad building process. One of the biggest challenges experts in this field have identified is the relative lack of technical expertise when it comes to using computers to perform analysis. Many of the tools currently being developed require a great deal of knowledge about computer systems integration, database management, network protocols, storage arrangements, and hardware and software specifications. This knowledge is an atypical skillset for a historian, and experts in many other fields, who could take advantage of information gained through data analysis.[20]

Summary

The literature reviewed shows many challenges and opportunities for organizations and fields to make better use of the data they collect. Big Data will continue to be an emerging issue, characterized primarily by its volume, variety, and velocity. KDD is a well-developed methodology for extracting information from Big Data and data sets. Analytics have provided a means for organizations to put the information from KDD to use. Visualizations have improved the communication of the results from KDD and increased the speed at which analytics can assist the decision making process.

However, there are many opportunities within the existing methodologies to create standardized processes. There has been little consideration for the time which a user must spend becoming an expert on the tools and processes. The user must spend additional time performing analysis to obtain results. No generic automated support for data exploration or creating visualizations is available. Users must devise their own methods in spreadsheet and statistical analysis software to explore and understand data. Each graph or visual must be created through wizards and the creation requires the user to manually manipulate and select data. The author presents a methodology to allow data exploration to be performed quickly and easily. The methodology incorporates visualizations into data exploration and automates their creation to further simplify the task for the user.

CHAPTER 3

Methodology

Overview

This methodology is divided into three main phases: Data Acquisition, Data Profiling, and Data Summaries and Visualization. The phases gather data from a user and provide information to the user about a data set. The flowchart in Figure 2 provides a high-level view of the process. Each of these phases will be covered in detail, with incremental steps.

The purpose of the methodology is to perform data exploration and use the information gained from the data and visualizations as part of a larger emphasis on analytics. In the field of Big Data, many different approaches have been used to tackle the problem of gaining information from large amounts of data. The author's process seeks to take a small amount of data and make the information in the data accessible extremely quickly. It provides the information in easily understandable formats, including charts and tabular summaries.

An implementation of this process should be a stand-alone software tool, requiring no background support systems or connections to existing databases or data warehouses. The interface should coincide with the average user's existing knowledge and basic use of common spreadsheet and personal computer software. The benefit of these requirements for implementation is to minimize any potential learning curve while exposing users to the potential benefits of analytics.

The process is not designed as a replacement for any existing tool or software. It is focused only on the data exploration phase of data mining, with the application of visualization to shortcut the more complex and lengthy analytics process. It fills the void for analysis tools between spreadsheet software and software requiring the use of more than one system.

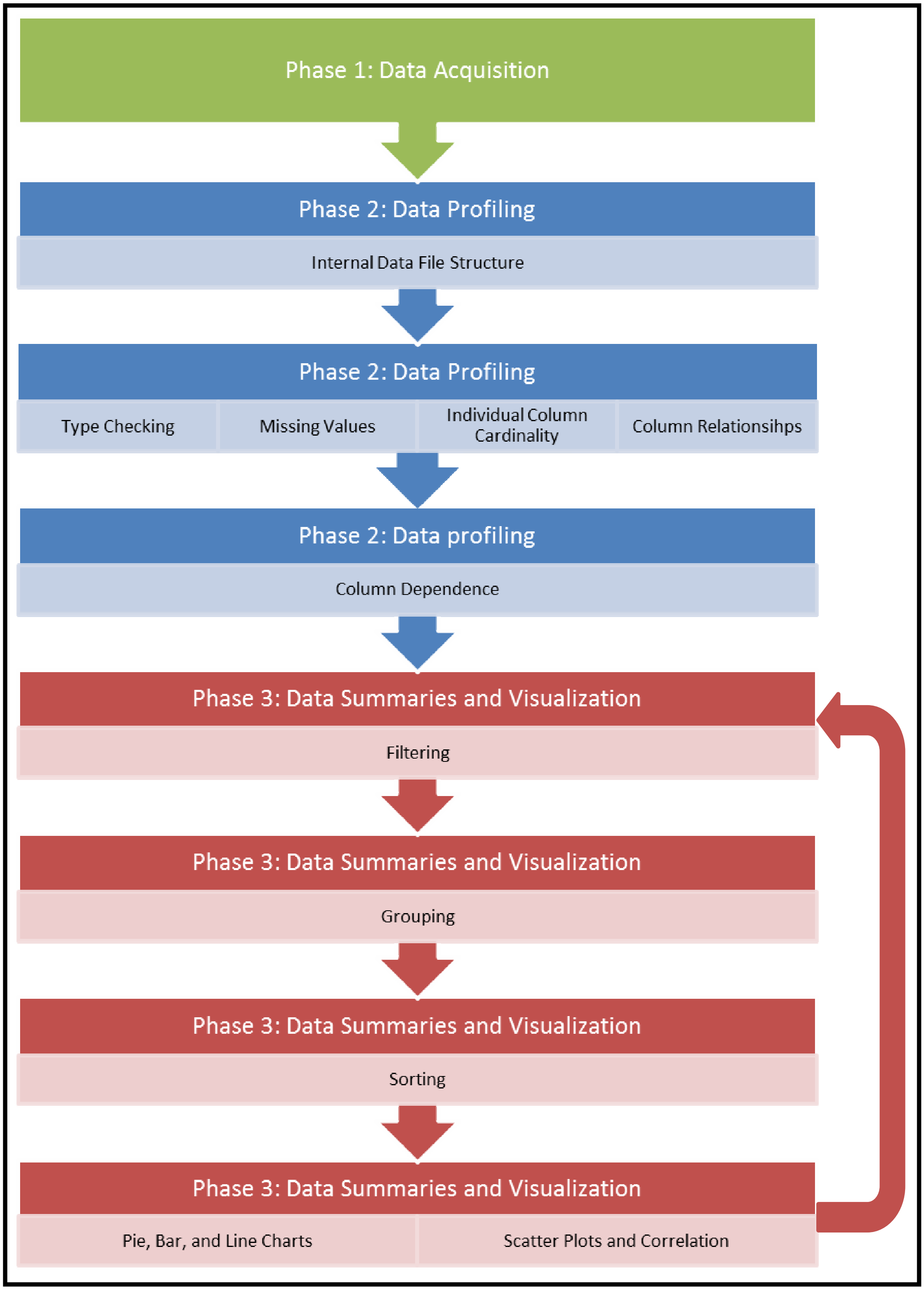

Phase 1: Data Acquisition
Phase 2: Data Profiling
Internal Data File Structure
Phase 2: Data Profiling
Type Checking
Missing Values
Individual Column Cardinality
Column Relationsihps
Phase 2: Data profiling
Column Dependence
Phase 3: Data Summaries and Visualization
Filtering
Phase 3: Data Summaries and Visualization
Grouping
Phase 3: Data Summaries and Visualization
Sorting
Phase 3: Data Summaries and Visualization
Pie, Bar, and Line Charts
Scatter Plots and Correlation

Fig. 2. Methodology Process Flowchart

Phase 1: Data Acquisition

The user must provide data for analysis. This is a non-trivial step in an application designed to be stand-alone. The data must also be limited in both rows and columns, since implementations of this application would be running on a personal computer. An entire data set of Big Data greatly exceeds the capabilities of a personal computer. A sample of data from a larger Big Data warehouse could be extracted for use. The data must also be in an easily consumable format. A plain text document, such as a comma-delimited or tab-delimited file, works well in many environments and eliminates excess storage space from padding fields. The user should be able to view the data in spreadsheet software, if needed, to update fields or view specific data points.

Phase 2: Data Profiling

Once the data is available, several steps can be taken to create a profile of the data and map out the contents. This phase of the process can be highly automated. User input and feedback will be required to get a complete profile of the data and to avoid potential performance losses from computationally expensive or inefficient tasks.

Internal Data File Structure

An important step in the process of profiling any data set is to understand the internal structure of the data that is provided. For a text based file such as a comma separated values spreadsheet, this includes the delimiting characters between fields, the delimiting character between fields containing the field delimiter, and the presence of field names as headers on each column of data. Many databases and data warehouses store field names separately from the data and queries append this information to the data. The user must be prompted to provide information about the internal structure before any automated analysis can begin.

Headers

An automated means to identify whether the data has field names or headers is to check to see that the first row is the only row with different data types. However, in the rare case that column headers are not text and are specialized types (times, dates, and numbers), the headers could be incorrectly skipped. Similarly, text headers on columns containing text data will not be differentiated.

The easiest way to determine whether the data has headers is to prompt the user to provide that information and display a visual cue showing the difference in appearance between the data with embedded headers or the data without headers.

Type Checking

In any data set of rows and columns, the types of data contained within each column are important to determine what analysis tools can be used to gain information. Type checking is also a valuable data cleanliness tool to verify there are no individual data points within a column that do not conform to the type format of the rest of the column. Anomalies such as these can prevent the column from being analyzed using type-specific techniques and, in some cases, prevent any further analysis.

Type checking can be performed through a variety of means. The most common method is to parse each data point within a column looking for a specific number, currency, date, percentage, or other non-text format. After each individual cell has been reviewed for type, the entire column can be cast to the least specific type that was discovered within the column. This conversion can be used for all further data profiling and analyses.

Missing Values

Missing values within data sets are a common occurrence. Missing values can be created during data collection or during data retrieval, or may be an indicator of a larger data cleanliness issue. These values can also have impacts on the value of statistical analyses performed on any data sets.

There are several methods for correcting missing values. Imputation replaces the missing value using the exact value from a complete row of data which is considered most similar to the row with the missing value. Interpolation estimates the value using the entire column of values to calculate the missing value. Pairwise deletion removes the entire row of data which contains a blank value in any column. The process proposed in this methodology only seeks to alert the user about the quantities of missing values.

Determining whether a value is missing requires assumptions to be made about what a null or missing value appears to be for different columns of data. Many datasets have specific default values used in place of no data (e.g. N/A for not available), which can make it harder to automatically detect missing values. The easiest missing values to report and detect are completely blank values in cells a given row and column.

Individual Column Cardinality

Cardinality is the number of unique values contained within a set. The number of unique values in each column is important to understand whether potential patterns can exist. If every column has the same number of unique values as the total number of rows, it is much less likely that patterns within the data will emerge. Uniqueness is itself a pattern within a column of data. If the number of unique values is one or more orders of magnitude smaller than the total number of values, this could provide an insight leading to further analysis.

Column Relationships

There may be a relationship of values between columns. These relationships show hierarchies, implied and hidden, between various columns. The relationships can identify data cleanliness issues when implied relationships do not exist where they are expected to. Relationships between columns can be used in data modeling and data storage. Relationships allow the effective reduction of the amount of data collected and the amount of storage required. These relationships exist in three different types as follows.

One to One

Each unique value in one column has only one corresponding value in another column for any row of data. These two columns might contain similar information, one column may be an abbreviation of another column, or there is a strictly linear relationship between the two columns.

One to Many

Each unique value in one column has many corresponding values in another column for various rows of data, but each value in that other column only has one corresponding value in the original column in any row of data. This is most commonly observed in stored hierarchy of information.

Many to Many

Each value in one column has many corresponding values in another column for various rows of data. This case may not represent an actual relationship but there may be subsets within a many to many relationship which can provide useful information.

Column Dependence

In any data set to be analyzed, there are independent and dependent variables. In scientific experiments these are also often referred to as factors and responses, respectively. The dependent variables are measured over time and not a characteristic of a process or experiment. Independent variables can be varied or fixed, but are a characteristics of a process or experiment.

The data type of columns can be used to grasp whether a column is independent or dependent. Typically, numerical measurements are dependent variables, while text fields are independent variables. However, with the emerging variety of data, text fields may be responses in an experiment, and independent text variables may be coded numerically for storage in a database. Similar to the case with column headers, the easiest solution to avoid possible confusion and errors is to prompt the user to identify which columns contain independent and dependent variables. Some intelligence can be used to automatically suggest the dependence of columns containing certain types of data, but users should be able to modify these suggestions.

Phase 3: Data Summaries and Visualization

For detailed review of the data contained within the selected file, a variety of different graphical displays showing the summarized contents can be used to quickly convey information to the user. In addition to the graphical displays, a tabular display of the supporting data should be readily available for viewing.

Summarizing the data requires the data to be filtered, grouped, and sorted so the information is presented in an easy to use format. These three steps are also difficult to perform manually across data sets in traditional analysis tools. Automating these steps provides further benefit to the analyst. The steps are performed in this order to improve the performance. Filtering the data reduces the total data which must be grouped, grouping the data compresses the information further, and sorting the information from the grouped data is also a smaller task.

Filtering

When dealing with quite large sets of data, the ability to filter data based on specific criteria is beneficial to determine more specific information. By default, all of the data will be analyzed and displayed in all graphs and figures. Filters allow focused analysis to be performed on subsets of the data provided. An outlier or trend may be noticeable after reviewing a tabular summary or visualization. Filtering to the specific data gives more detailed view of the information.

Grouping

The data set will have previously identified independent and dependent variables contained within columns. The user should be able to focus on one independent variable with respect to dependent variables. The resulting dependent variable data is grouped or binned by values of the independent variable. Descriptive statistics about each group of data can be calculated to provide information which is not readily accessible.

Sorting

Once information about the dependent variable groups for each independent variable value has been calculated, these values should be ordered by the independent variable in a logical order. The original data is not guaranteed to be in any specific order and may not be logical. Grouping the data may change the overall order of the original data but does not guarantee a particular order. After sorting, any tabular or visual representations of the information generated from the data will then have a discernible organization and order.

Pie, Bar, and Line Charts

The sorted information is used to create multiple types of charts to display the information. Because the data has an unknown and generic context, a specific graph that is most appropriate for the information cannot be created. Instead, several different charts are created and the user can switch between the charts appropriately. The three charts selected for the basic display of information include a pie chart, a bar chart, and a line chart.

Correlation and Scatter Plots

Using two dependent variables at a time instead of a single variable, scatter plots are constructed using the same grouping by a single independent variable, but for both dependent variables. The correlation coefficient between the two dependent variables is calculated to determine if there is a relationship between these values. In a tabular view, statistics about both variables is provided with the correlation coefficient.

A scatter plot with one dependent variable on the horizontal axis and one dependent variable on the vertical axis shows the correlation visually. Since multiple groups of data can be expected, the ability to filter the data to a single group or a subset of the total data is still beneficial.

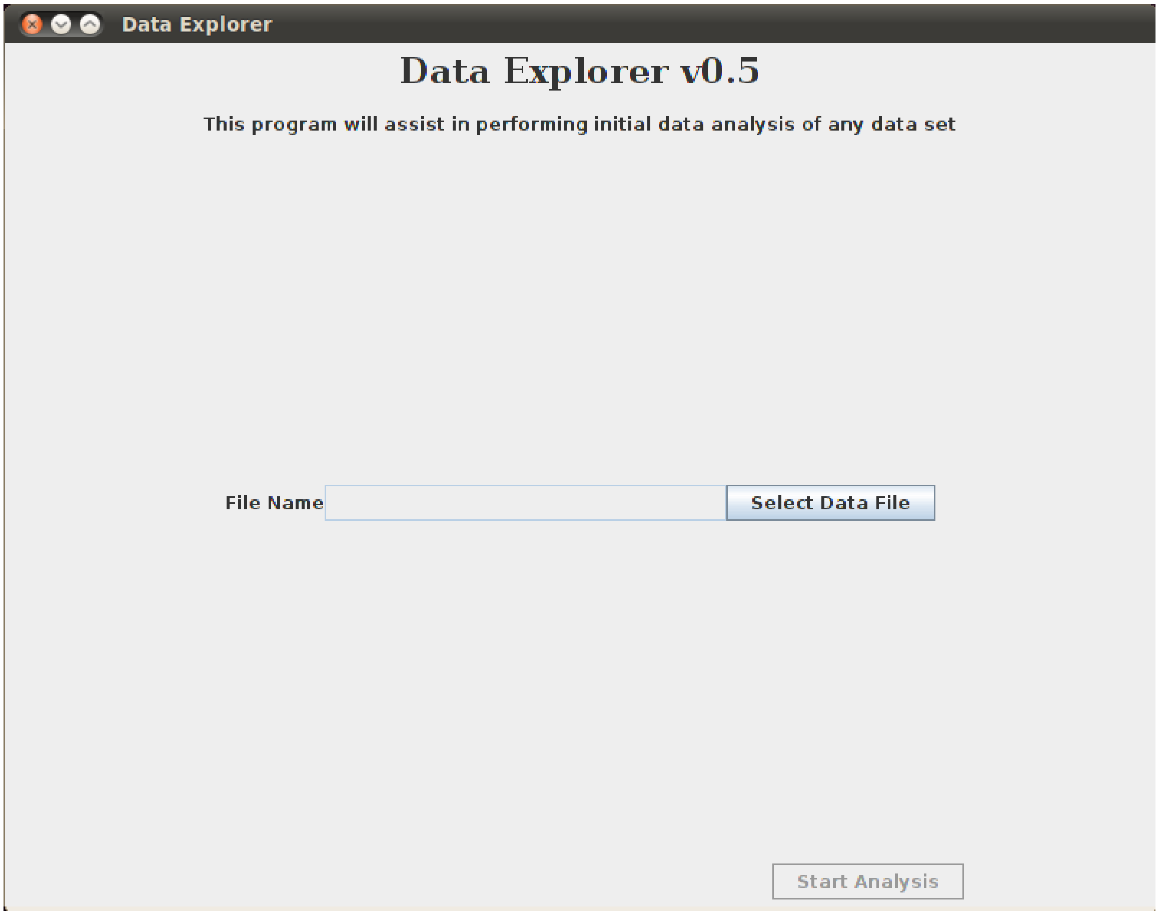
Implementation

A Java prototype application implementing the majority of this methodology was created for testing. The prototype relies on the Java Swing user interface components, charts from the JFreeChart Java package, and Java Utility data structures including hash tables, sets, array lists, arrays, and vectors. Figure 3 shows the application layout and window.

Limitations

The implementation in Java runs through the Java Virtual Machine, in memory on a personal computer. This restricts the total footprint of the program resources and the data to be within the computer's available memory. The implementation handles two data types, text and numerical values represented as integers or precision floating points.

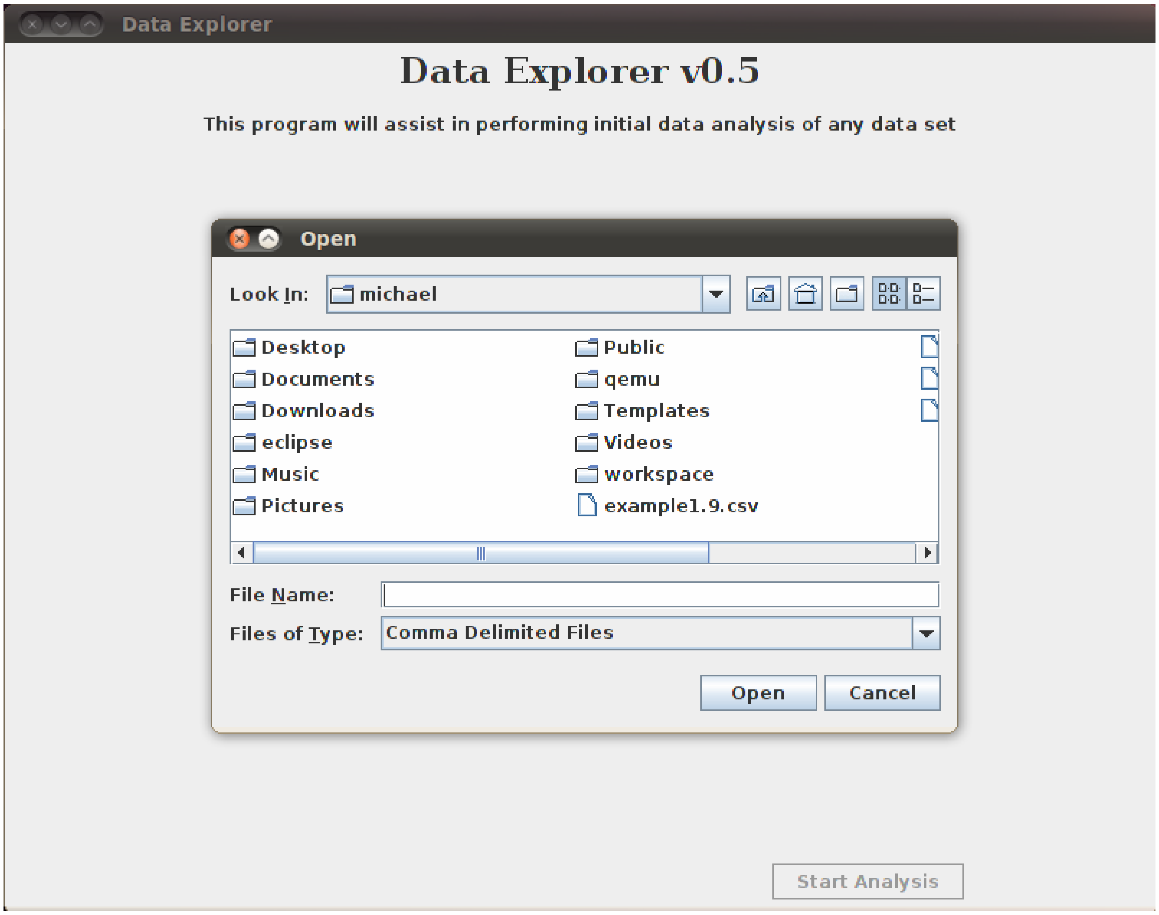
The implementation also uses a file selection tool and accepts only comma separated value format for data files as seen in Figure 4.



Data Explorer v0.5

This program will assist in performing initial data analysis of any data set

Fig. 3. Welcome Screen



Data Explorer v0.5

This program will assist in performing initial data analysis of any data set

Look In: michael

File Name:

Files of Type: Comma Delimited Files

Fig. 4. File Selection

CHAPTER 4

Analysis and Case Studies

The following are two case studies performed by the author using the proposed methodology. Each case provides challenges and interesting questions which cannot be easily answered by viewing the raw data in a spreadsheet. The two cases demonstrate a broader range of applicability across more other fields. These applications will be discussed in depth for each case.

Case 1: Correlates of War Project

The Correlates of War Project is an ongoing effort by historians, economists, and in some cases environmentalists, to gauge the "power" of nations around the world. The definition of power used by the project is "the ability of a nation to exercise and resist influence – [as] a function of many factors, among them the nation's material capabilities". The most significant output from this project is the National Material Capabilities (NMC) data set. The data set contains annual values for total population, urban population, iron and steel production, energy consumption, military personnel, and military expenditure of all state members. The available data set currently contains data from 1816-2007. The widely-used Composite Index of National Capability (CINC) index is based on these six variables and included in the data set.[6][32][13]

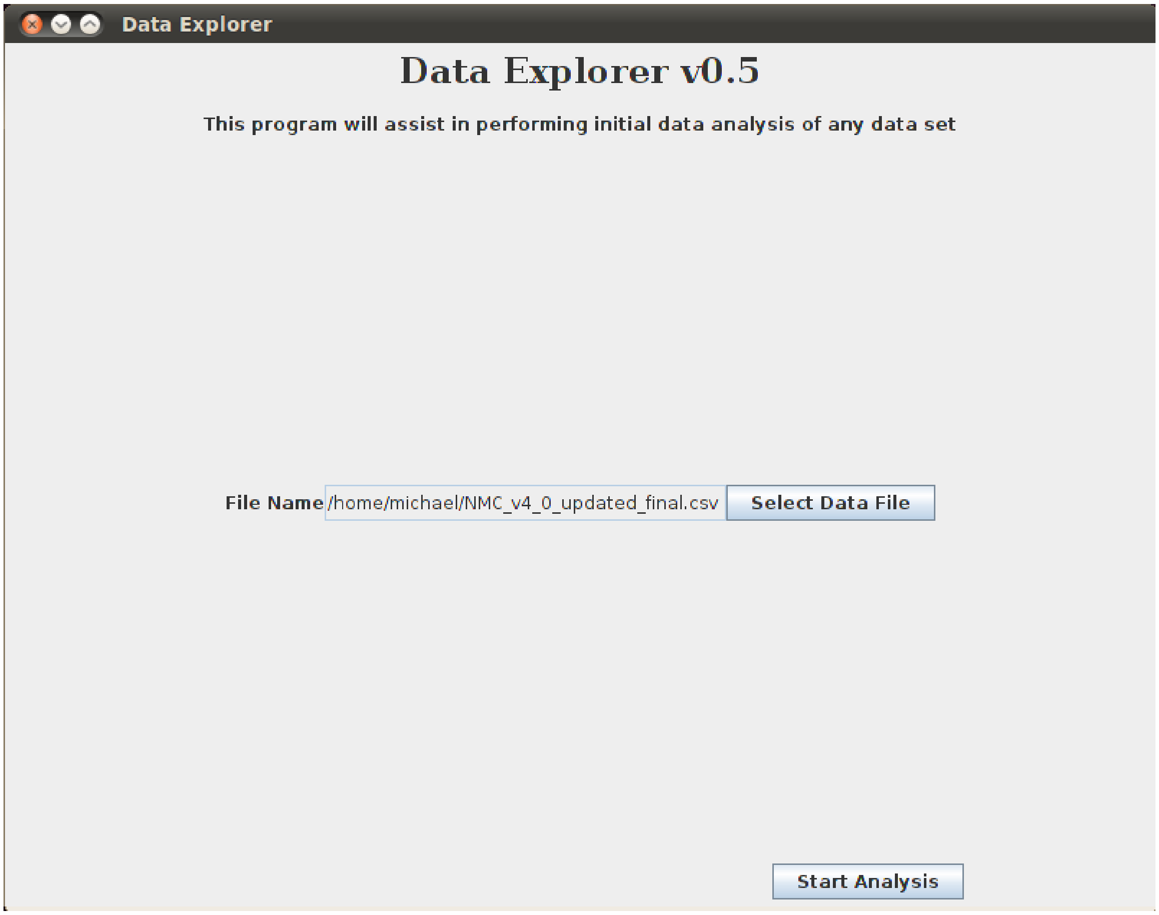
This data set was selected because little analysis has been performed in the past on this data and the measurements recorded could be directly relevant to a wide variety of applications. Total population, urban population, and military personnel are all measurements of human resources. Energy consumption and military expenditures are both measurements of costs, both direct and indirect. Iron and steel production is a measurement of outputs. Human resources, direct and indirect costs, and outputs are applicable to any organizations focused on managing personnel, financial information, and meeting goals.

A revised version of the NMC data set incorporating additional data, including the continents, country names, and decades, was used to provide more detailed information. The additional data was derived from the original data set which included country codes and years.

The following information, graphs, and data were obtained using the prototype implementation of the methodology. Figures 5 through 25 are screenshots taken directly from the prototype implementation and can be used as a guide to replicate the analysis.

Generating these graphs and information by hand would take a considerable amount of time for an individual analyst. Using the implementation, simple button clicks and toggle settings allow for summaries of various categories of data to be generated within seconds.

Phase 1: Data Acquisition



Data Explorer v0.5

This program will assist in performing initial data analysis of any data set

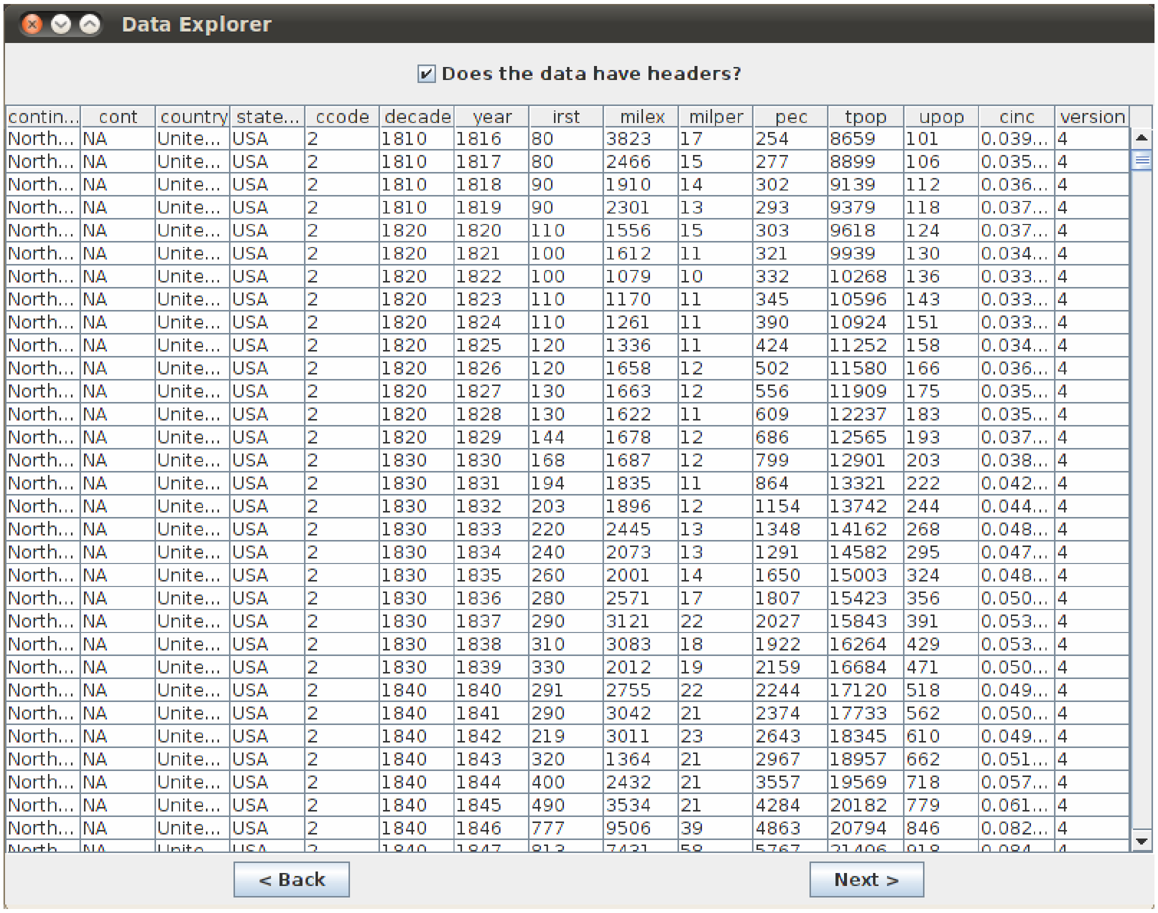
File Name /home/michael/NMC_v4_0_updated_final.csv

Fig. 5. NMC Data File Selected

Once a data file has been selected through the pop up window, the path to the file appears in the file name box. As seen in Figure 5, the Start Analysis button is no longer disabled and the user can now begin the process. For comparison, see Figure 3 to view the disabled version of the button.

Phase 2: Data Profiling

Internal Data File Structure



Data Explorer

☑ Does the data have headers?

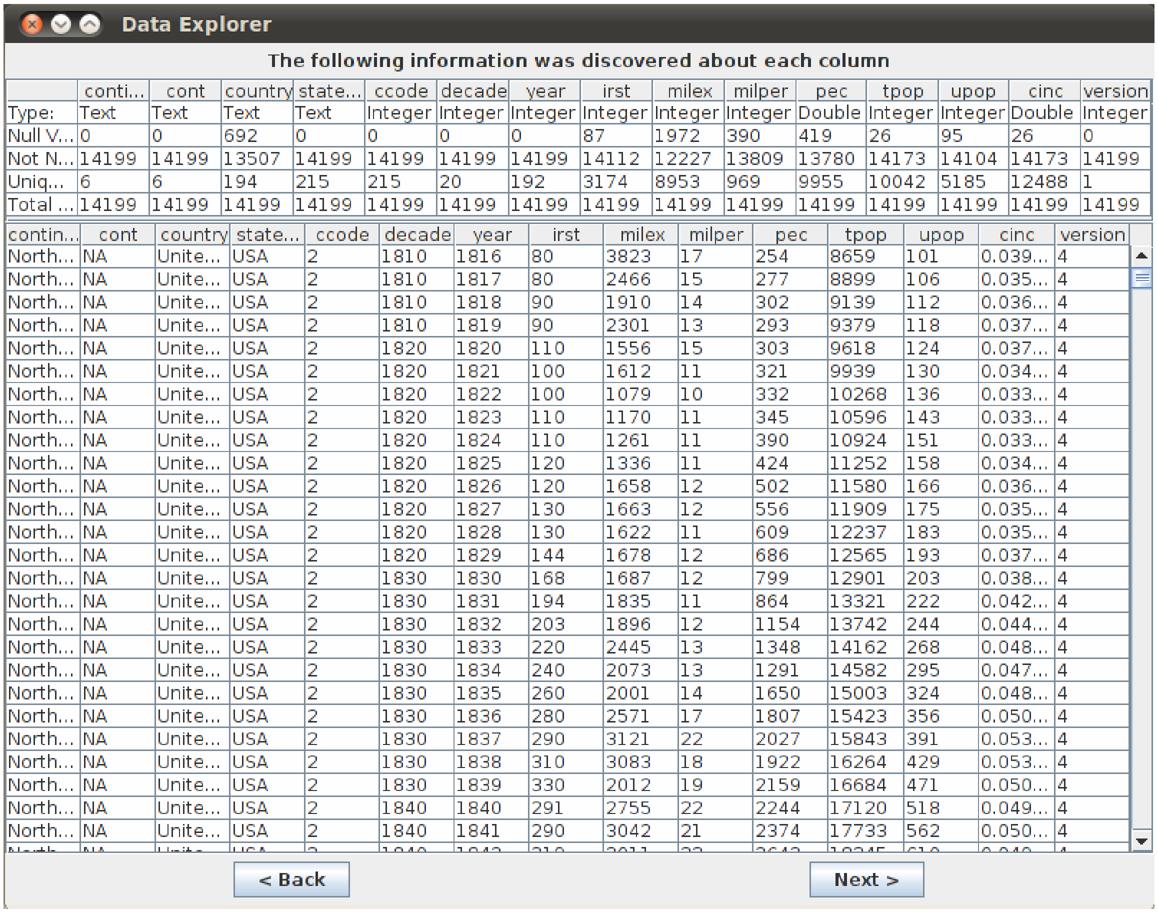
contin...	cont	country	state...	ccode	decade	year	irst	milex	milper	pec	tpop	upop	cinc	version
North...	NA	Unite...	USA	2	1810	1816	80	3823	17	254	8659	101	0.039...	4
North...	NA	Unite...	USA	2	1810	1817	80	2466	15	277	8899	106	0.035...	4
North...	NA	Unite...	USA	2	1810	1818	90	1910	14	302	9139	112	0.036...	4
North...	NA	Unite...	USA	2	1810	1819	90	2301	13	293	9379	118	0.037...	4
North...	NA	Unite...	USA	2	1820	1820	110	1556	15	303	9618	124	0.037...	4
North...	NA	Unite...	USA	2	1820	1821	100	1612	11	321	9939	130	0.034...	4
North...	NA	Unite...	USA	2	1820	1822	100	1079	10	332	10268	136	0.033...	4
North...	NA	Unite...	USA	2	1820	1823	110	1170	11	345	10596	143	0.033...	4
North...	NA	Unite...	USA	2	1820	1824	110	1261	11	390	10924	151	0.033...	4
North...	NA	Unite...	USA	2	1820	1825	120	1336	11	424	11252	158	0.034...	4
North...	NA	Unite...	USA	2	1820	1826	120	1658	12	502	11580	166	0.036...	4
North...	NA	Unite...	USA	2	1820	1827	130	1663	12	556	11909	175	0.035...	4
North...	NA	Unite...	USA	2	1820	1828	130	1622	11	609	12237	183	0.035...	4
North...	NA	Unite...	USA	2	1820	1829	144	1678	12	686	12565	193	0.037...	4
North...	NA	Unite...	USA	2	1830	1830	168	1687	12	799	12901	203	0.038...	4
North...	NA	Unite...	USA	2	1830	1831	194	1835	11	864	13321	222	0.042...	4
North...	NA	Unite...	USA	2	1830	1832	203	1896	12	1154	13742	244	0.044...	4
North...	NA	Unite...	USA	2	1830	1833	220	2445	13	1348	14162	268	0.048...	4
North...	NA	Unite...	USA	2	1830	1834	240	2073	13	1291	14582	295	0.047...	4
North...	NA	Unite...	USA	2	1830	1835	260	2001	14	1650	15003	324	0.048...	4
North...	NA	Unite...	USA	2	1830	1836	280	2571	17	1807	15423	356	0.050...	4
North...	NA	Unite...	USA	2	1830	1837	290	3121	22	2027	15843	391	0.053...	4
North...	NA	Unite...	USA	2	1830	1838	310	3083	18	1922	16264	429	0.053...	4
North...	NA	Unite...	USA	2	1830	1839	330	2012	19	2159	16684	471	0.050...	4
North...	NA	Unite...	USA	2	1840	1840	291	2755	22	2244	17120	518	0.049...	4
North...	NA	Unite...	USA	2	1840	1841	290	3042	21	2374	17733	562	0.050...	4
North...	NA	Unite...	USA	2	1840	1842	219	3011	23	2643	18345	610	0.049...	4
North...	NA	Unite...	USA	2	1840	1843	320	1364	21	2967	18957	662	0.051...	4
North...	NA	Unite...	USA	2	1840	1844	400	2432	21	3557	19569	718	0.057...	4
North...	NA	Unite...	USA	2	1840	1845	490	3534	21	4284	20182	779	0.061...	4
North...	NA	Unite...	USA	2	1840	1846	777	9506	39	4863	20794	846	0.082...	4
North...	NA	Unite...	USA	2	1840	1847	813	7431	58	5767	21406	918	0.084...	4

Fig. 6. NMC Headers Verified

The tool has a toggle check box for the user to identify the presence of headers. If the user selects no headers, the columns are numbered in order with default names. The NMC data set is conveniently provided as a comma separated value spreadsheet. There are also headers at the top of the sheet which contain acronyms to discern the contents of the column. The headers shown at the top of Figure 6 correspond to the following:

- continent – the name of each continent
- cont – the two letter abbreviation for each continent
- country – the name of each country
- stateabb – the three letter abbreviation for each country
- ccode – the Correlates of War country code
- decade – the decade
- year – the year
- irst – iron and steel production in thousands of tons
- milex – military expenditures
- milper – military personnel in thousands
- pec – primary energy consumption in thousands of coal-ton equivalents
- tpop – total population in thousands
- upop – urban population in thousands, (city population over 100,000)
- cinc – Composite Index of National Capability (CINC) score
- version – data set version number

Type Checking, Missing Values, and Individual Column Cardinality



The following information was discovered about each column

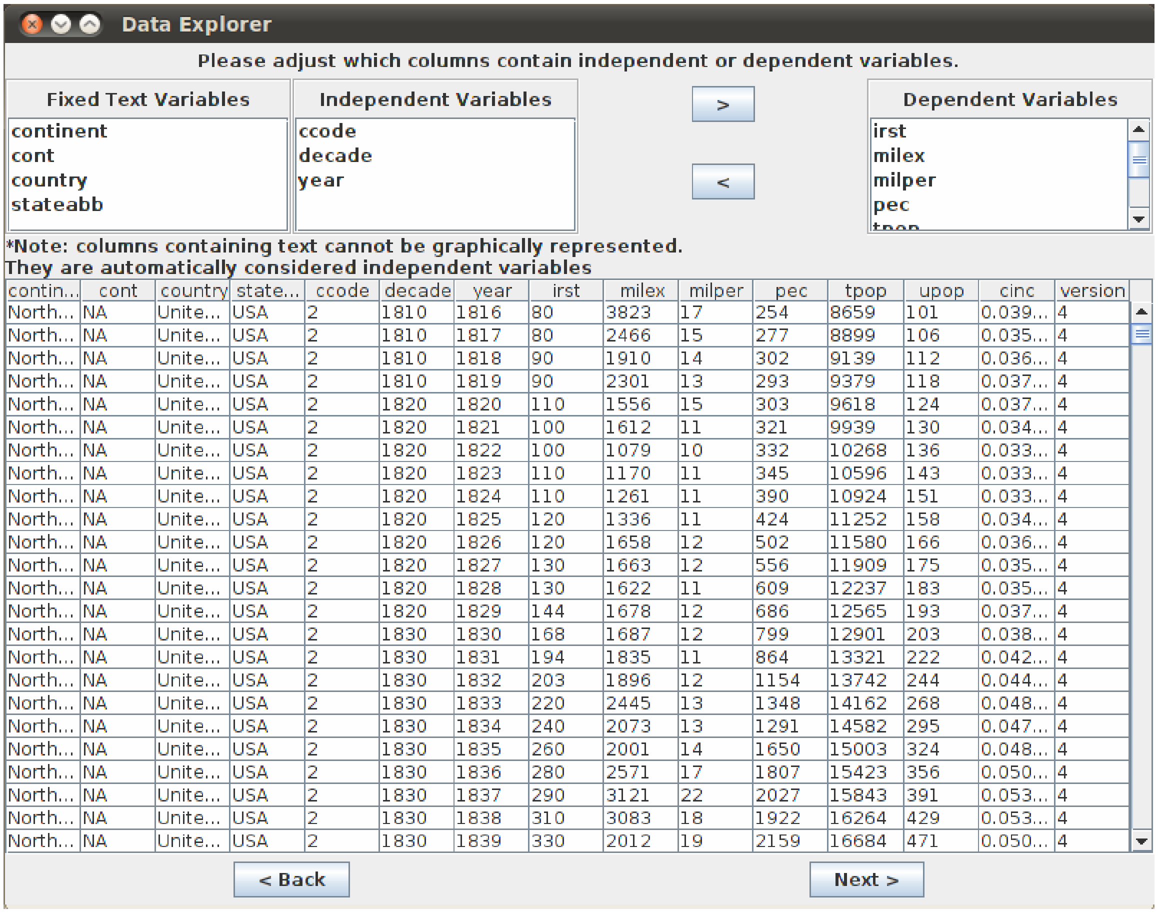
	conti...	cont	country	state...	ccode	decade	year	irst	milex	milper	pec	tpop	upop	cinc	version
Type:	Text	Text	Text	Text	Integer	Integer	Integer	Integer	Integer	Integer	Double	Integer	Integer	Double	Integer
Null V...	0	0	692	0	0	0	0	87	1972	390	419	26	95	26	0
Not N...	14199	14199	13507	14199	14199	14199	14199	14112	12227	13809	13780	14173	14104	14173	14199
Uniq...	6	6	194	215	215	20	192	3174	8953	969	9955	10042	5185	12488	1
Total ...	14199	14199	14199	14199	14199	14199	14199	14199	14199	14199	14199	14199	14199	14199	14199

contin...	cont	country	state...	ccode	decade	year	irst	milex	milper	pec	tpop	upop	cinc	version
North...	NA	Unite...	USA	2	1810	1816	80	3823	17	254	8659	101	0.039...	4
North...	NA	Unite...	USA	2	1810	1817	80	2466	15	277	8899	106	0.035...	4
North...	NA	Unite...	USA	2	1810	1818	90	1910	14	302	9139	112	0.036...	4
North...	NA	Unite...	USA	2	1810	1819	90	2301	13	293	9379	118	0.037...	4
North...	NA	Unite...	USA	2	1820	1820	110	1556	15	303	9618	124	0.037...	4
North...	NA	Unite...	USA	2	1820	1821	100	1612	11	321	9939	130	0.034...	4
North...	NA	Unite...	USA	2	1820	1822	100	1079	10	332	10268	136	0.033...	4
North...	NA	Unite...	USA	2	1820	1823	110	1170	11	345	10596	143	0.033...	4
North...	NA	Unite...	USA	2	1820	1824	110	1261	11	390	10924	151	0.033...	4
North...	NA	Unite...	USA	2	1820	1825	120	1336	11	424	11252	158	0.034...	4
North...	NA	Unite...	USA	2	1820	1826	120	1658	12	502	11580	166	0.036...	4
North...	NA	Unite...	USA	2	1820	1827	130	1663	12	556	11909	175	0.035...	4
North...	NA	Unite...	USA	2	1820	1828	130	1622	11	609	12237	183	0.035...	4
North...	NA	Unite...	USA	2	1820	1829	144	1678	12	686	12565	193	0.037...	4
North...	NA	Unite...	USA	2	1830	1830	168	1687	12	799	12901	203	0.038...	4
North...	NA	Unite...	USA	2	1830	1831	194	1835	11	864	13321	222	0.042...	4
North...	NA	Unite...	USA	2	1830	1832	203	1896	12	1154	13742	244	0.044...	4
North...	NA	Unite...	USA	2	1830	1833	220	2445	13	1348	14162	268	0.048...	4
North...	NA	Unite...	USA	2	1830	1834	240	2073	13	1291	14582	295	0.047...	4
North...	NA	Unite...	USA	2	1830	1835	260	2001	14	1650	15003	324	0.048...	4
North...	NA	Unite...	USA	2	1830	1836	280	2571	17	1807	15423	356	0.050...	4
North...	NA	Unite...	USA	2	1830	1837	290	3121	22	2027	15843	391	0.053...	4
North...	NA	Unite...	USA	2	1830	1838	310	3083	18	1922	16264	429	0.053...	4
North...	NA	Unite...	USA	2	1830	1839	330	2012	19	2159	16684	471	0.050...	4
North...	NA	Unite...	USA	2	1840	1840	291	2755	22	2244	17120	518	0.049...	4
North...	NA	Unite...	USA	2	1840	1841	290	3042	21	2374	17733	562	0.050...	4

Fig. 7. NMC Column Summaries

The tool has identified a relatively large number of null values for the military expenditures column, over 10% of the total data. Null values such as these are ignored in any statistics to prevent skew from unknown data. The version number identifies the source of the data on every row with the same single value in every row as seen in Figure 7. The user can click the next button to move on at any point in time.

Column Dependence



Please adjust which columns contain independent or dependent variables.

Fixed Text Variables
continent
cont
country
stateabb

Independent Variables
ccode
decade
year

Dependent Variables
irst
milex
milper
pec
tpop

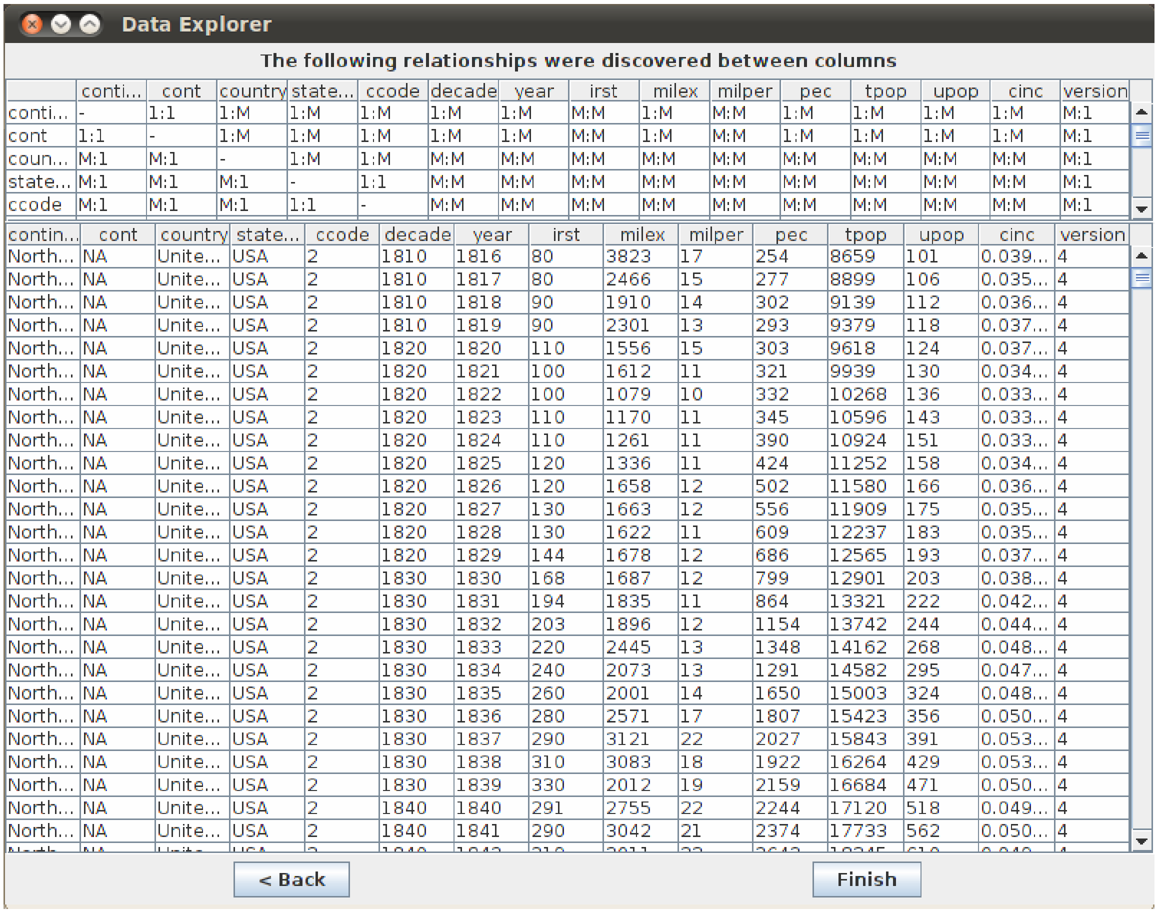
*Note: columns containing text cannot be graphically represented.
They are automatically considered independent variables

contin...	cont	country	state...	ccode	decade	year	irst	milex	milper	pec	tpop	upop	cinc	version
North...	NA	Unite...	USA	2	1810	1816	80	3823	17	254	8659	101	0.039...	4
North...	NA	Unite...	USA	2	1810	1817	80	2466	15	277	8899	106	0.035...	4
North...	NA	Unite...	USA	2	1810	1818	90	1910	14	302	9139	112	0.036...	4
North...	NA	Unite...	USA	2	1810	1819	90	2301	13	293	9379	118	0.037...	4
North...	NA	Unite...	USA	2	1820	1820	110	1556	15	303	9618	124	0.037...	4
North...	NA	Unite...	USA	2	1820	1821	100	1612	11	321	9939	130	0.034...	4
North...	NA	Unite...	USA	2	1820	1822	100	1079	10	332	10268	136	0.033...	4
North...	NA	Unite...	USA	2	1820	1823	110	1170	11	345	10596	143	0.033...	4
North...	NA	Unite...	USA	2	1820	1824	110	1261	11	390	10924	151	0.033...	4
North...	NA	Unite...	USA	2	1820	1825	120	1336	11	424	11252	158	0.034...	4
North...	NA	Unite...	USA	2	1820	1826	120	1658	12	502	11580	166	0.036...	4
North...	NA	Unite...	USA	2	1820	1827	130	1663	12	556	11909	175	0.035...	4
North...	NA	Unite...	USA	2	1820	1828	130	1622	11	609	12237	183	0.035...	4
North...	NA	Unite...	USA	2	1820	1829	144	1678	12	686	12565	193	0.037...	4
North...	NA	Unite...	USA	2	1830	1830	168	1687	12	799	12901	203	0.038...	4
North...	NA	Unite...	USA	2	1830	1831	194	1835	11	864	13321	222	0.042...	4
North...	NA	Unite...	USA	2	1830	1832	203	1896	12	1154	13742	244	0.044...	4
North...	NA	Unite...	USA	2	1830	1833	220	2445	13	1348	14162	268	0.048...	4
North...	NA	Unite...	USA	2	1830	1834	240	2073	13	1291	14582	295	0.047...	4
North...	NA	Unite...	USA	2	1830	1835	260	2001	14	1650	15003	324	0.048...	4
North...	NA	Unite...	USA	2	1830	1836	280	2571	17	1807	15423	356	0.050...	4
North...	NA	Unite...	USA	2	1830	1837	290	3121	22	2027	15843	391	0.053...	4
North...	NA	Unite...	USA	2	1830	1838	310	3083	18	1922	16264	429	0.053...	4
North...	NA	Unite...	USA	2	1830	1839	330	2012	19	2159	16684	471	0.050...	4

Fig. 8. NMC Independent & Dependent Variable Selection

In the NMC data set, the country code, decade, and year are all integer values and could be considered dependent variables. As shown in Figure 8, selection of the items and the toggle arrows allow movement between the independent and dependent variable lists. Text variables are restricted from consideration as dependent variables because they cannot be represented graphically. The user is also prevented from considering all variables either independent or dependent.

Column Relationships



The following relationships were discovered between columns

	conti...	cont	country	state...	ccode	decade	year	irst	milex	milper	pec	tpop	upop	cinc	version
conti...	-	1:1	1:M	1:M	1:M	1:M	1:M	M:M	1:M	M:M	1:M	1:M	1:M	1:M	M:1
cont	1:1	-	1:M	1:M	1:M	1:M	1:M	M:M	1:M	M:M	1:M	1:M	1:M	1:M	M:1
coun...	M:1	M:1	-	1:M	1:M	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:1
state...	M:1	M:1	M:1	-	1:1	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:1
ccode	M:1	M:1	M:1	1:1	-	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:M	M:1

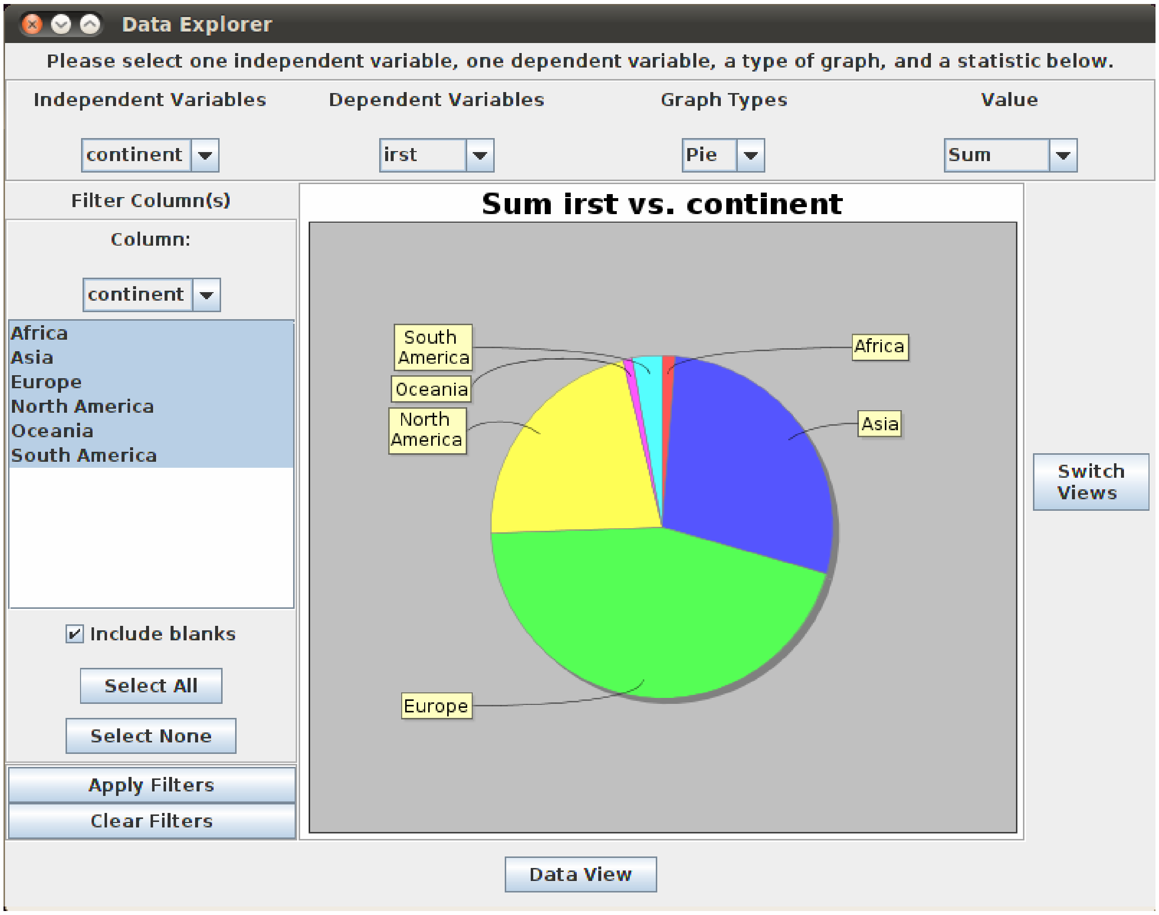
contin...	cont	country	state...	ccode	decade	year	irst	milex	milper	pec	tpop	upop	cinc	version
North...	NA	Unite...	USA	2	1810	1816	80	3823	17	254	8659	101	0.039...	4
North...	NA	Unite...	USA	2	1810	1817	80	2466	15	277	8899	106	0.035...	4
North...	NA	Unite...	USA	2	1810	1818	90	1910	14	302	9139	112	0.036...	4
North...	NA	Unite...	USA	2	1810	1819	90	2301	13	293	9379	118	0.037...	4
North...	NA	Unite...	USA	2	1820	1820	110	1556	15	303	9618	124	0.037...	4
North...	NA	Unite...	USA	2	1820	1821	100	1612	11	321	9939	130	0.034...	4
North...	NA	Unite...	USA	2	1820	1822	100	1079	10	332	10268	136	0.033...	4
North...	NA	Unite...	USA	2	1820	1823	110	1170	11	345	10596	143	0.033...	4
North...	NA	Unite...	USA	2	1820	1824	110	1261	11	390	10924	151	0.033...	4
North...	NA	Unite...	USA	2	1820	1825	120	1336	11	424	11252	158	0.034...	4
North...	NA	Unite...	USA	2	1820	1826	120	1658	12	502	11580	166	0.036...	4
North...	NA	Unite...	USA	2	1820	1827	130	1663	12	556	11909	175	0.035...	4
North...	NA	Unite...	USA	2	1820	1828	130	1622	11	609	12237	183	0.035...	4
North...	NA	Unite...	USA	2	1820	1829	144	1678	12	686	12565	193	0.037...	4
North...	NA	Unite...	USA	2	1830	1830	168	1687	12	799	12901	203	0.038...	4
North...	NA	Unite...	USA	2	1830	1831	194	1835	11	864	13321	222	0.042...	4
North...	NA	Unite...	USA	2	1830	1832	203	1896	12	1154	13742	244	0.044...	4
North...	NA	Unite...	USA	2	1830	1833	220	2445	13	1348	14162	268	0.048...	4
North...	NA	Unite...	USA	2	1830	1834	240	2073	13	1291	14582	295	0.047...	4
North...	NA	Unite...	USA	2	1830	1835	260	2001	14	1650	15003	324	0.048...	4
North...	NA	Unite...	USA	2	1830	1836	280	2571	17	1807	15423	356	0.050...	4
North...	NA	Unite...	USA	2	1830	1837	290	3121	22	2027	15843	391	0.053...	4
North...	NA	Unite...	USA	2	1830	1838	310	3083	18	1922	16264	429	0.053...	4
North...	NA	Unite...	USA	2	1830	1839	330	2012	19	2159	16684	471	0.050...	4
North...	NA	Unite...	USA	2	1840	1840	291	2755	22	2244	17120	518	0.049...	4
North...	NA	Unite...	USA	2	1840	1841	290	3042	21	2374	17733	562	0.050...	4

Fig. 9. NMC Column Relationships

The continent and cont columns contain values which can be treated as synonymous for the purposes of analysis, shown by the 1:1 in Figure 9, representing a one to one relationship. The ccode and stateabb columns also contain values which can be treated as synonymous. The decade and year columns show a clear hierarchy with a one to many relationship.

Phase 3: Data Summaries and Visualizations

Iron and Steel Production



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables	Dependent Variables	Graph Types	Value
continent	irst	Pie	Sum

Filter Column(s)

Column: continent

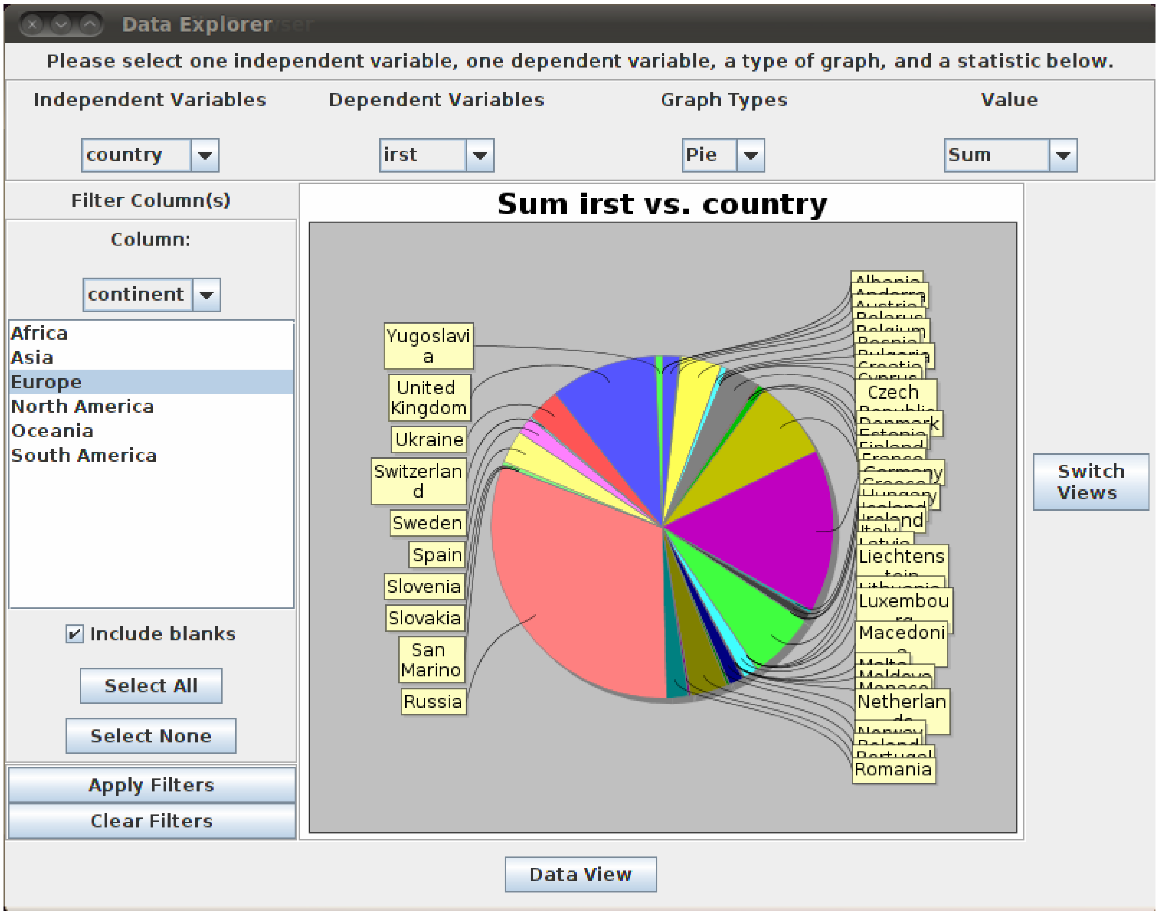
Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

Sum irst vs. continent

Fig. 10. Total Iron & Steel Production Per Continent – Pie Chart

Figure 10 displays the total recorded iron and steel production per continent from 1816-2007 in a pie chart. From this chart, further investigation into the continent producing the largest amount of iron and steel is a sensible choice. In Figure 11, the filter on the left hand side is adjusted to select only Europe and the independent variable is changed from continent to country.


Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: country

Dependent Variables: irst

Graph Types: Pie

Value: Sum

Filter Column(s)

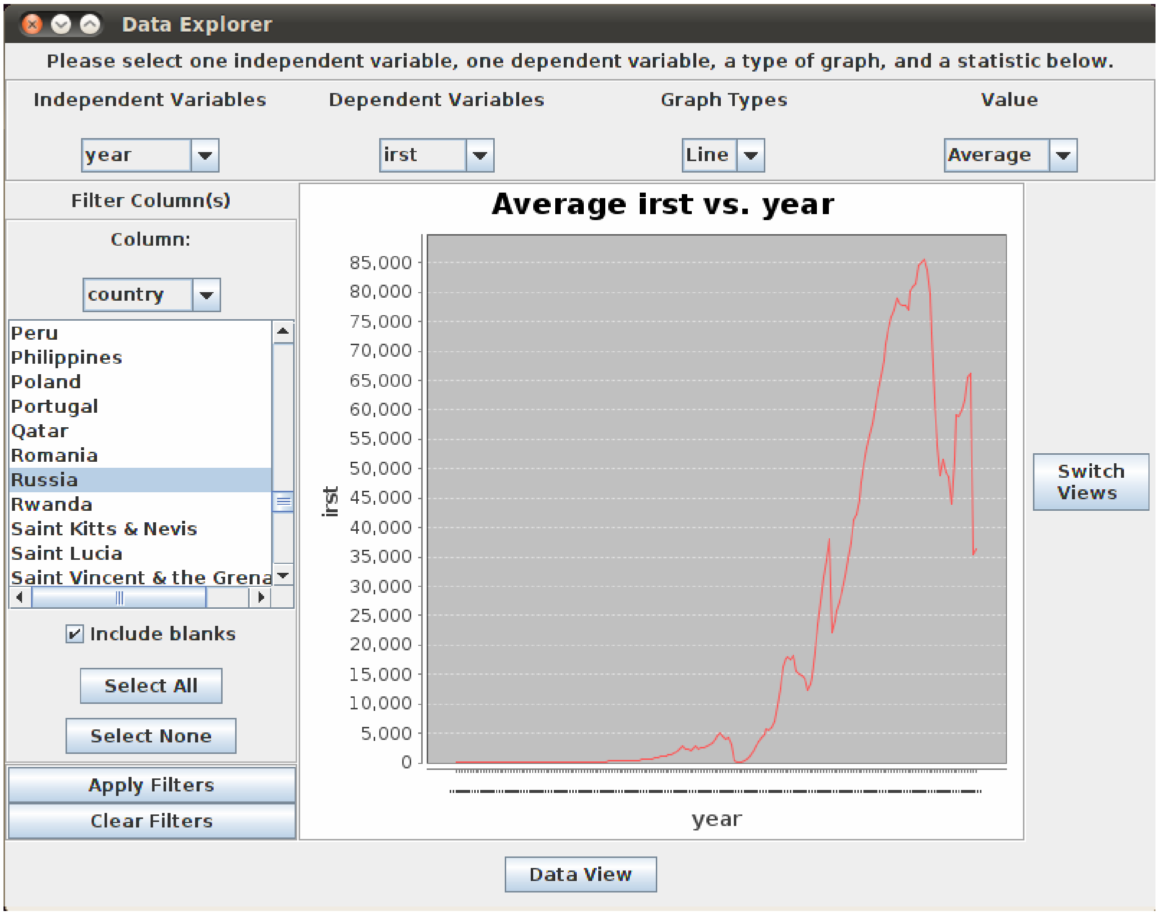
Column: continent

Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

Fig. 11. Total Iron & Steel Production Per Country in Europe – Pie Chart

In Figure 11, Russia (in salmon) is the largest producer of iron and steel over time in Europe, with Germany (in purple), the United Kingdom (in blue), France (in dark yellow), and Italy (in bright green) following closely. The data can be filtered to focus only on Russia, with the independent variable changed to year and the graph type changed to line as seen in Figure 12.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: year

Dependent Variables: irst

Graph Types: Line

Value: Average

Filter Column(s)

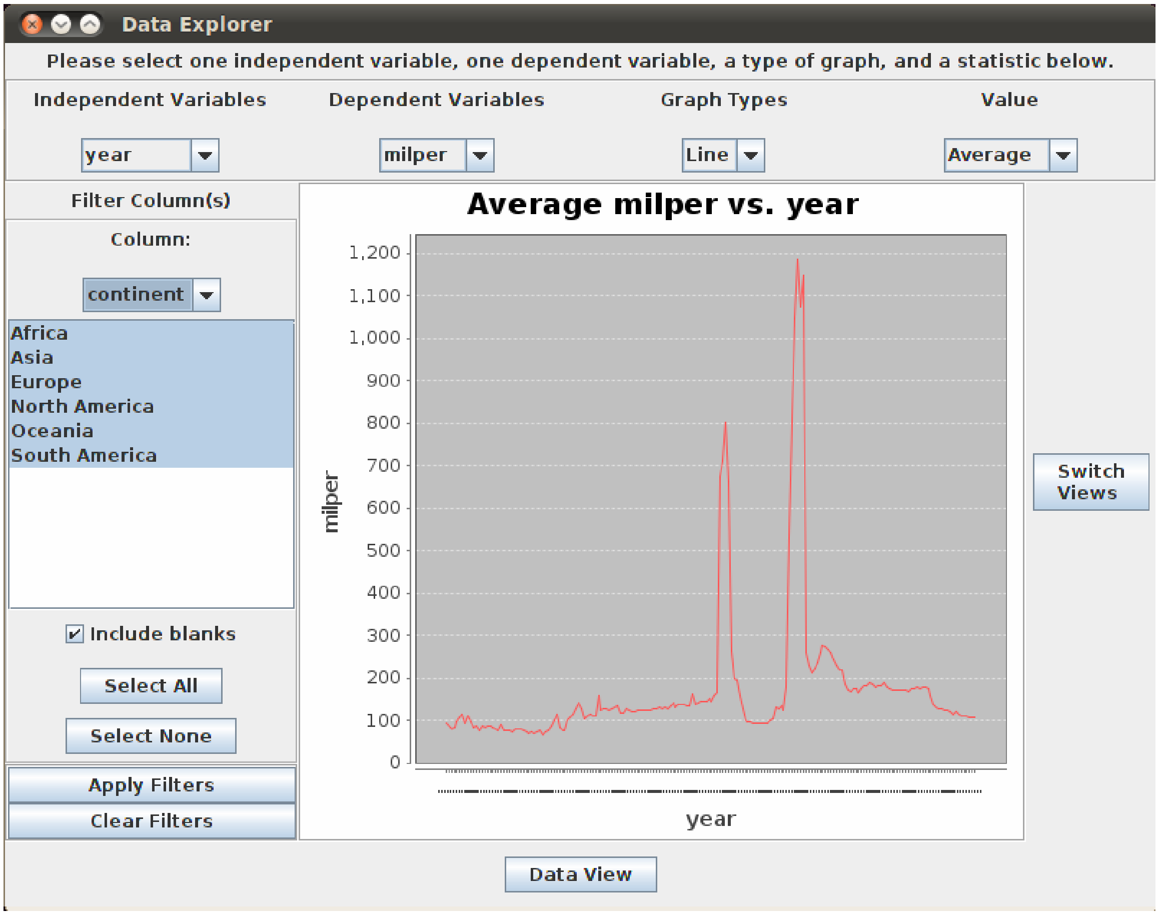
Column: country

☑ Include blanks

Fig. 12. Average Iron and Steel Production Per Year in Russia - Line Chart

The axis labels are unreadable because there are too many categories, the years 1816-2007. The data view is provided to clarify the missing labels and hovering over any points shows the specific value. Figure 12 contains the graph of Russian iron and steel production over time. It shows a large amount of growth and then a steep drop off. The largest drop is when the Soviet Union disbanded in the late 1980's and early 1990's.

Military Personnel



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables	Dependent Variables	Graph Types	Value
year	milper	Line	Average

Filter Column(s)

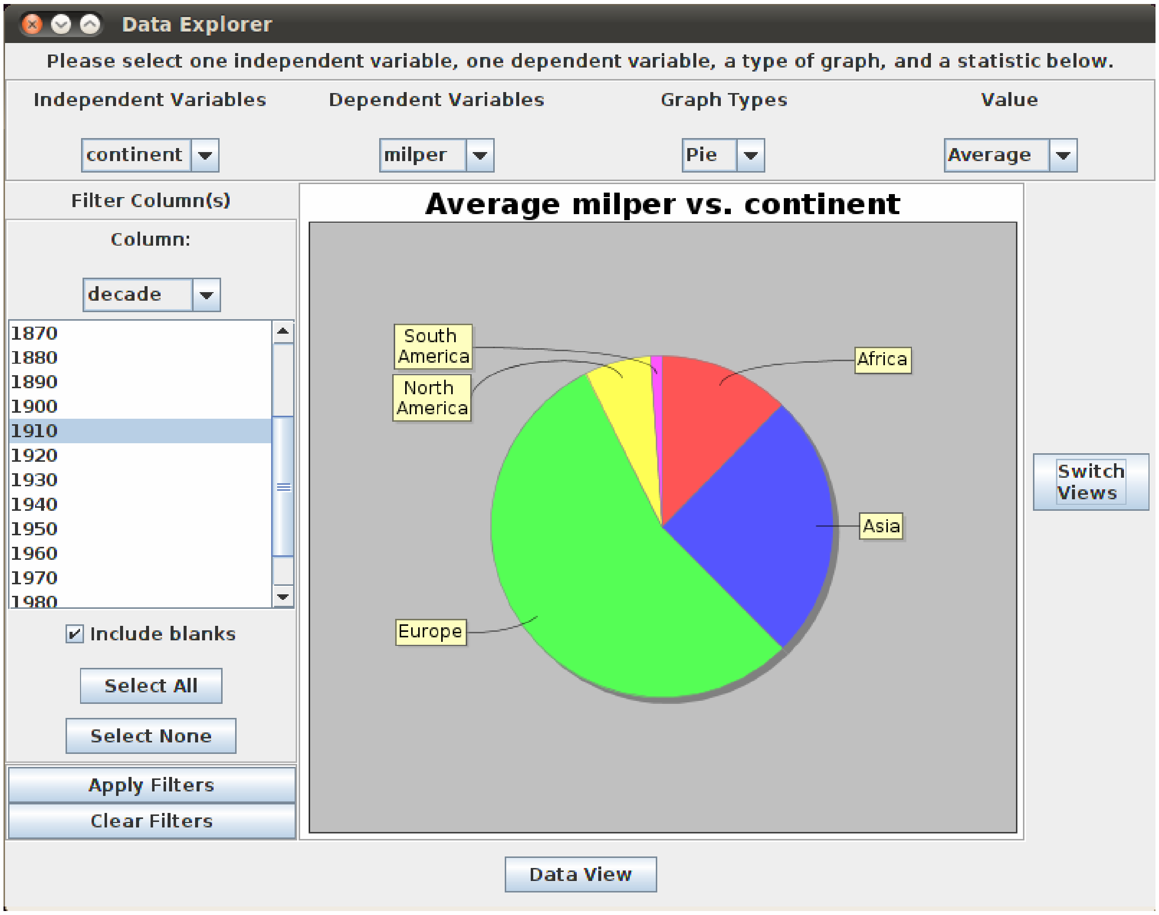
Column: continent

Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

Fig. 13. Average Military Personnel Per Year – Line Chart

Clearing all of the filters and viewing a chart of the average military personnel per year shows two significant events, World War I (WWI) and World War II (WWII) in Figure 13. Focusing on WWI, the decade can be filtered to view only the 1910's in a pie chart by continent to establish where the war took place as seen in Figure 14.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: continent

Dependent Variables: milper

Graph Types: Pie

Value: Average

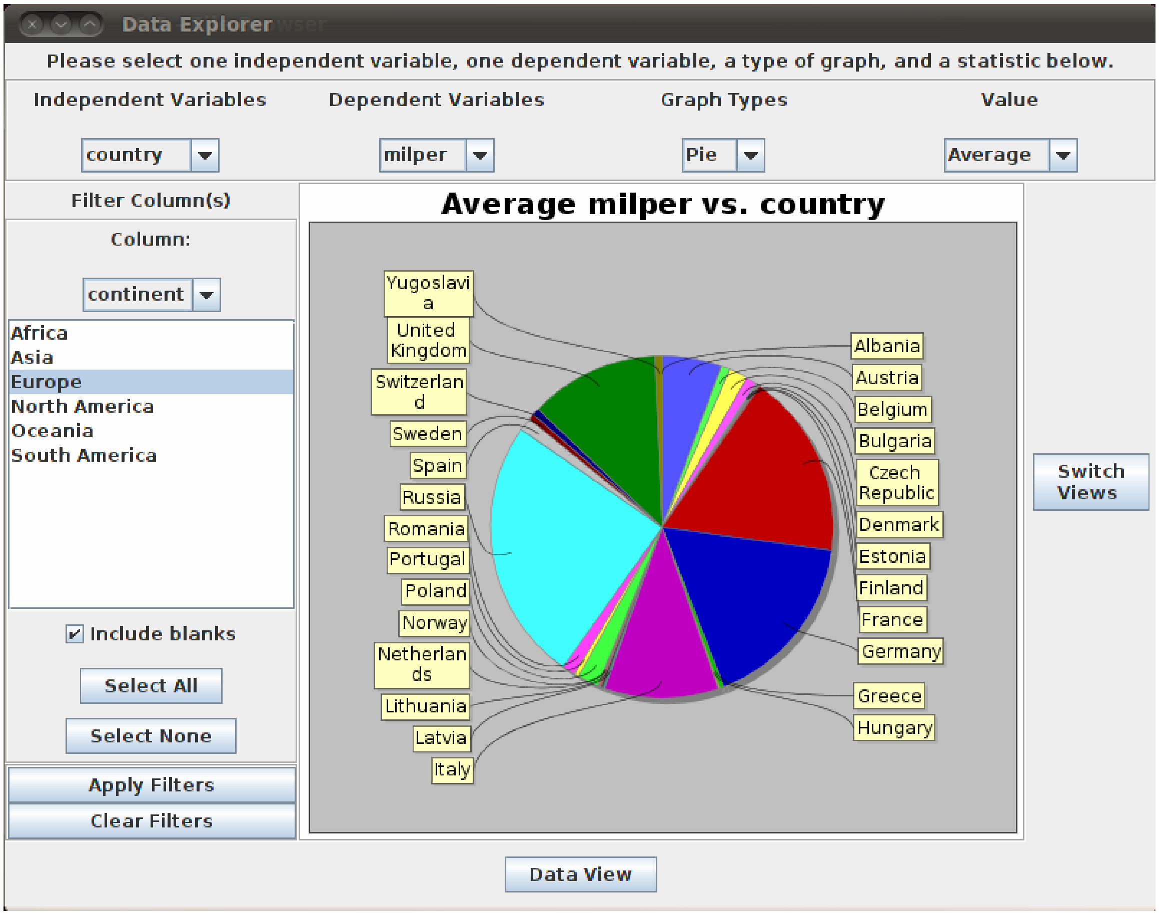
Filter Column(s)

Column: decade

☑ Include blanks

Fig. 14. Average Military Personnel Per Continent in the 1910's - Pie Chart

The majority of conflict during WWI occurred in Europe, which is readily observable from the chart in Figure 14. Changing the filter to focus only on Europe and the independent variable to country will show which countries were involved in the conflict as seen in Figure 15.



Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: country

Dependent Variables: milper

Graph Types: Pie

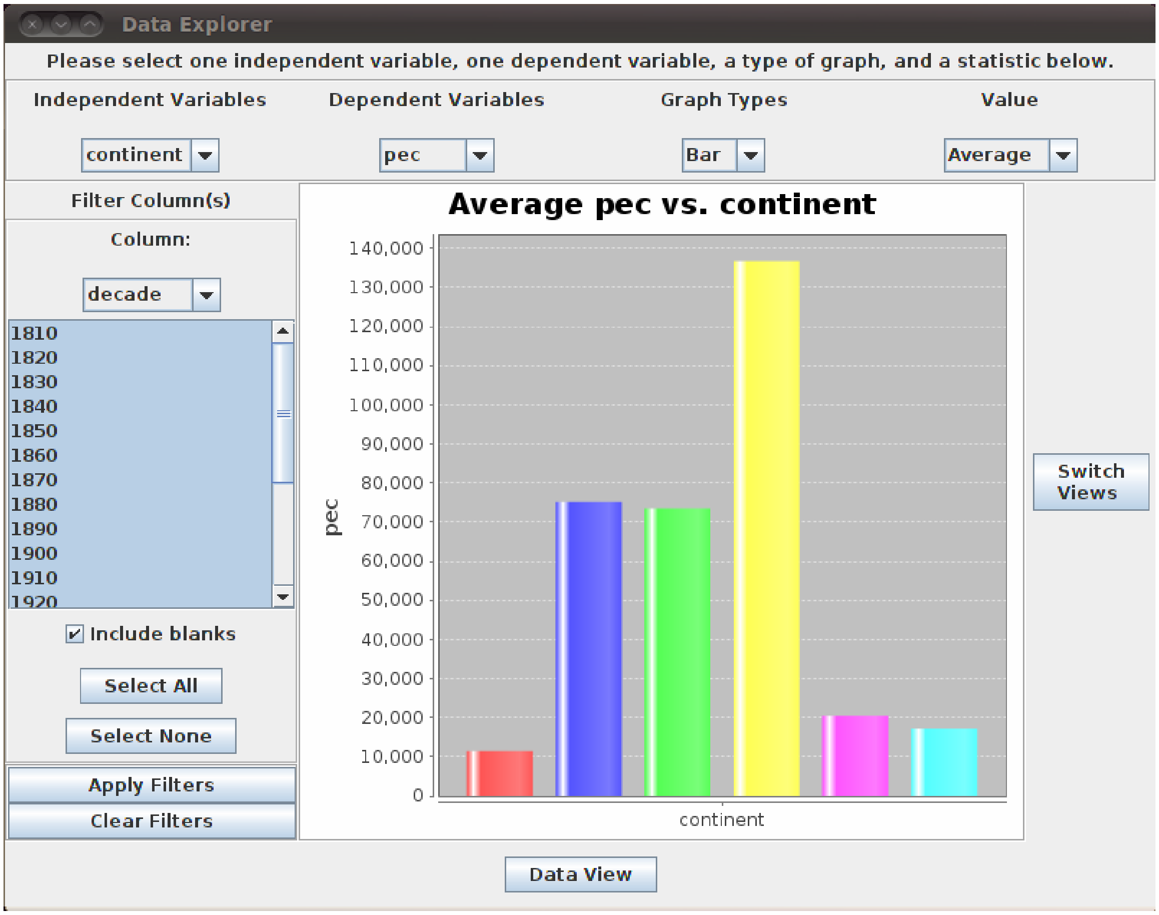
Value: Average

Column: continent

Fig. 15. Average Military Personnel Per Country in Europe in the 1910's - Pie Chart

From the chart in Figure 15, it appears that Austria, France, Germany, Italy, Russia, and the United Kingdom had significant involvement in the conflict. Austria is of note because it was not previously identified as a large producer of iron and steel, but still appears to have played a major role in WWI.

Primary Energy Consumption



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: continent

Dependent Variables: pec

Graph Types: Bar

Value: Average

Filter Column(s)

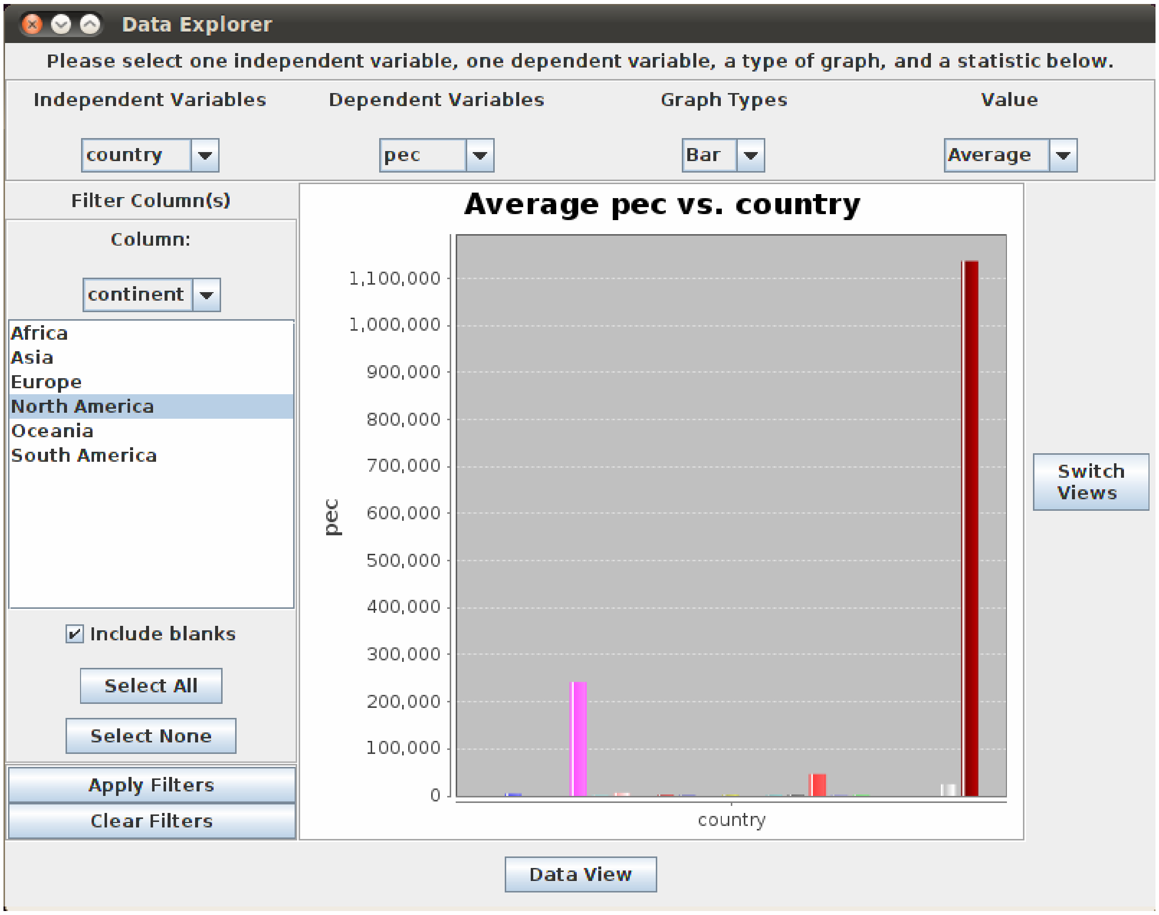
Column: decade

1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920

☑ Include blanks

Fig. 16. Average Primary Energy Consumption Per Continent - Bar Chart

In the bar chart, the categories are not visible but can be viewed on hover over. Again, clearing all of the filters and viewing the average primary energy consumption per continent, North America stands out as the largest consumer of thousands of tons of coal equivalent on average in Figure 16, represented by the yellow bar. The continent can be filtered to North America and the independent variable can be changed to country to get a more detailed view about the energy consumption as seen in Figure 17.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: country

Dependent Variables: pec

Graph Types: Bar

Value: Average

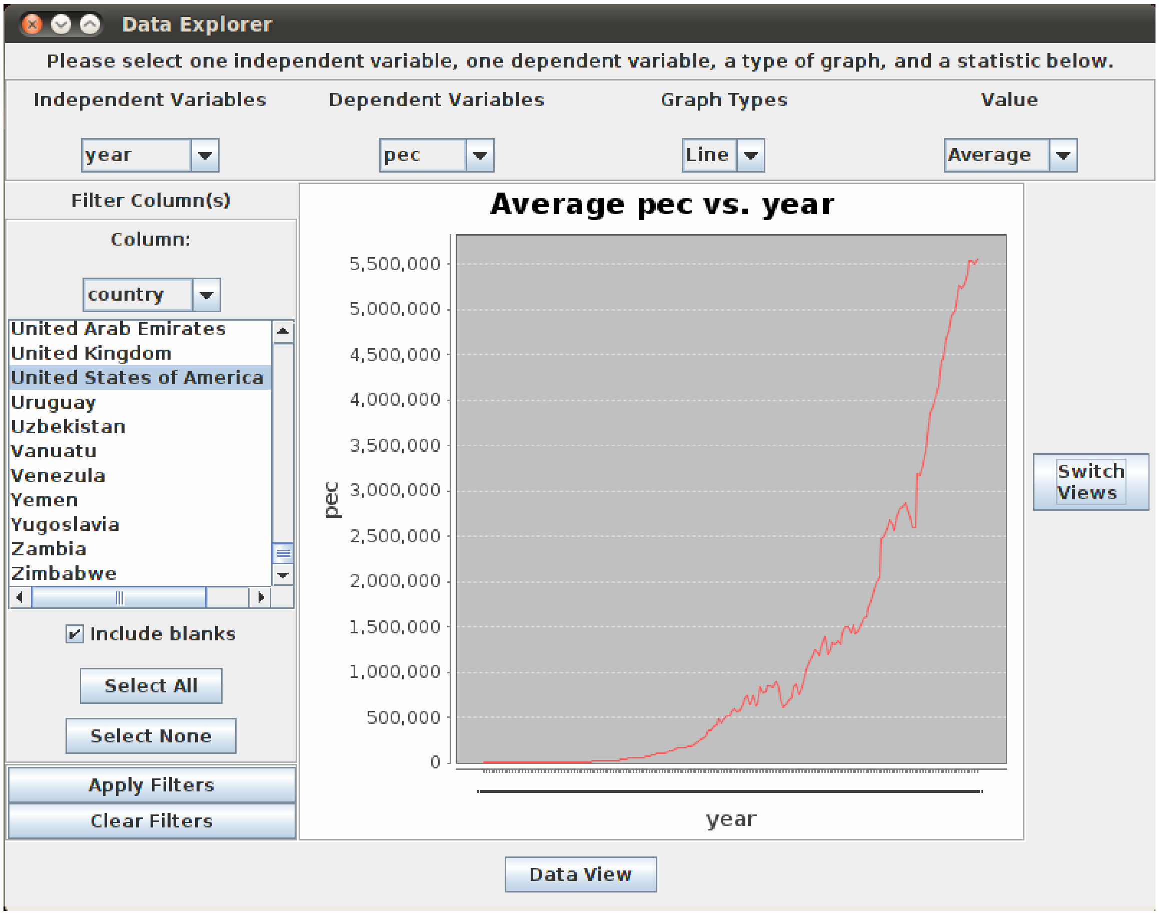
Filter Column(s)

Column: continent

☑ Include blanks

Fig. 17. Average Primary Energy Consumption Per Country in North America - Bar Chart

The United States is clearly the largest consumer of energy in North America in Figure 17, represented by the dark red bar. The other columns with a value of 0 indicate data which is incomplete: every country would have some energy consumption over the period of 1816-2007. A more detailed view of the United States consumption over time can be created by filtering the country to the United States, changing the independent variable to year, and changing the graph type to line as seen in Figure 18.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables	Dependent Variables	Graph Types	Value
year	pec	Line	Average

Filter Column(s)

Column: country

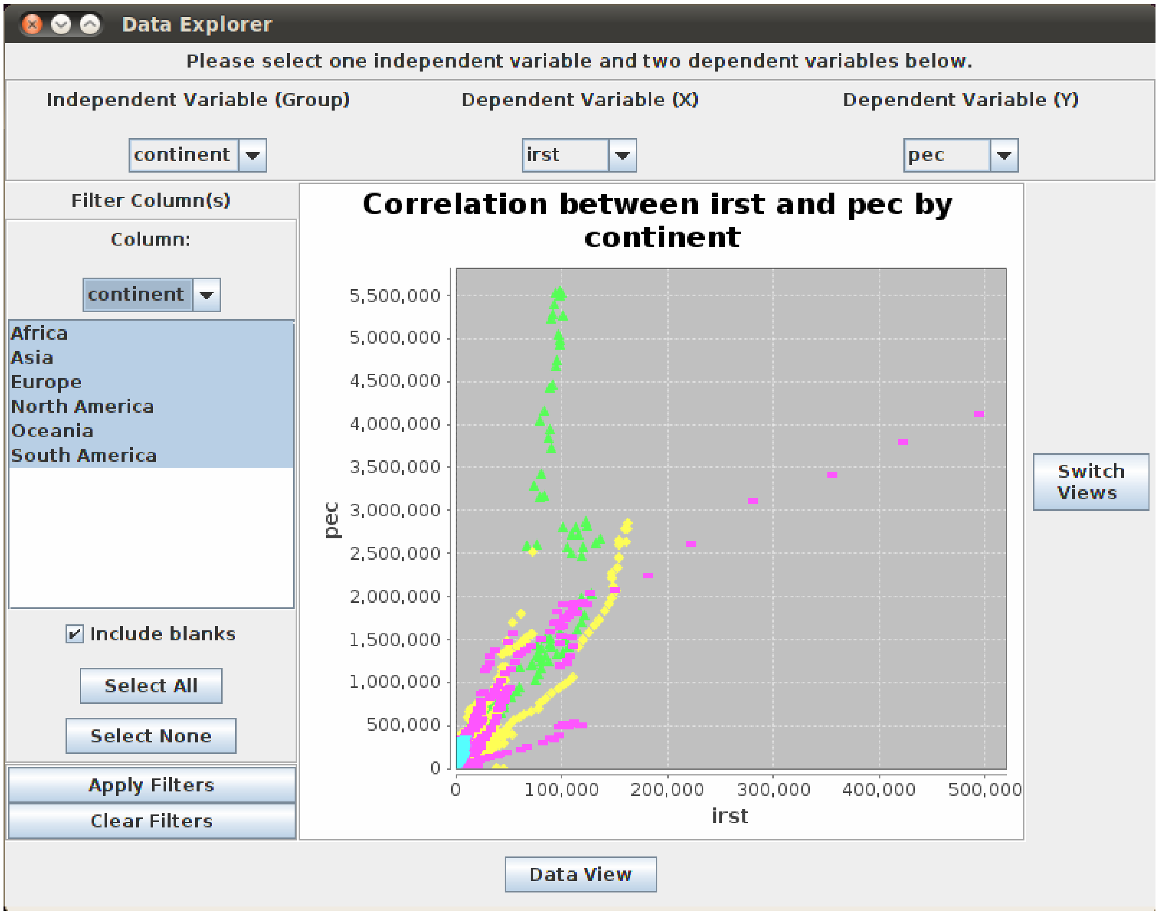
United Arab Emirates
United Kingdom
United States of America
Uruguay
Uzbekistan
Vanuatu
Venezula
Yemen
Yugoslavia
Zambia
Zimbabwe

☑ Include blanks

Fig. 18. Average Primary Energy Consumption Per Year in the United States - Line Chart

The energy consumption of the United States has grown significantly over time. The average across all of the data will be skewed by the most recent data in the tail of the distribution. There are some slight transitional periods seen in Figure 18 which cover the periods of WWI, the Great Depression, and WWII.

IRST and PEC Correlation by Continent



Data Explorer

Please select one independent variable and two dependent variables below.

Independent Variable (Group): continent

Dependent Variable (X): irst

Dependent Variable (Y): pec

Filter Column(s)

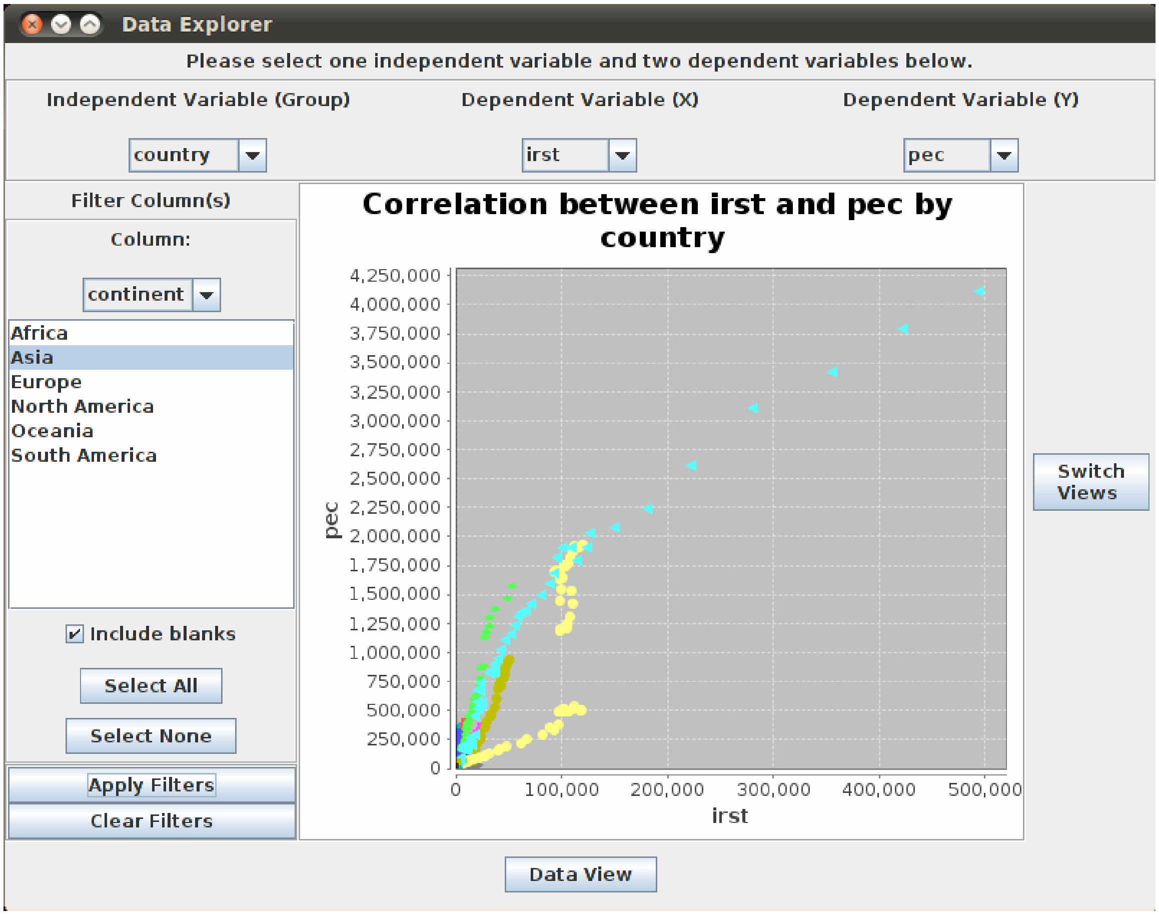
Column: continent

Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

Fig. 19. Primary Energy Consumption versus Iron and Steel Production by Continent - Scatter Plot

Clearing the filters and switching views, a scatter plot of two dependent variables can be created. In Figure 19, the primary energy consumption is plotted against the iron and steel production. The different series are for each continent, the independent variable. Focusing on Asia, represented in pink on Figure 19, the continent filter will be adjusted and the independent variable will be changed to country as seen in Figure 20.



Data Explorer

Please select one independent variable and two dependent variables below.

Independent Variable (Group): country

Dependent Variable (X): irst

Dependent Variable (Y): pec

Filter Column(s)

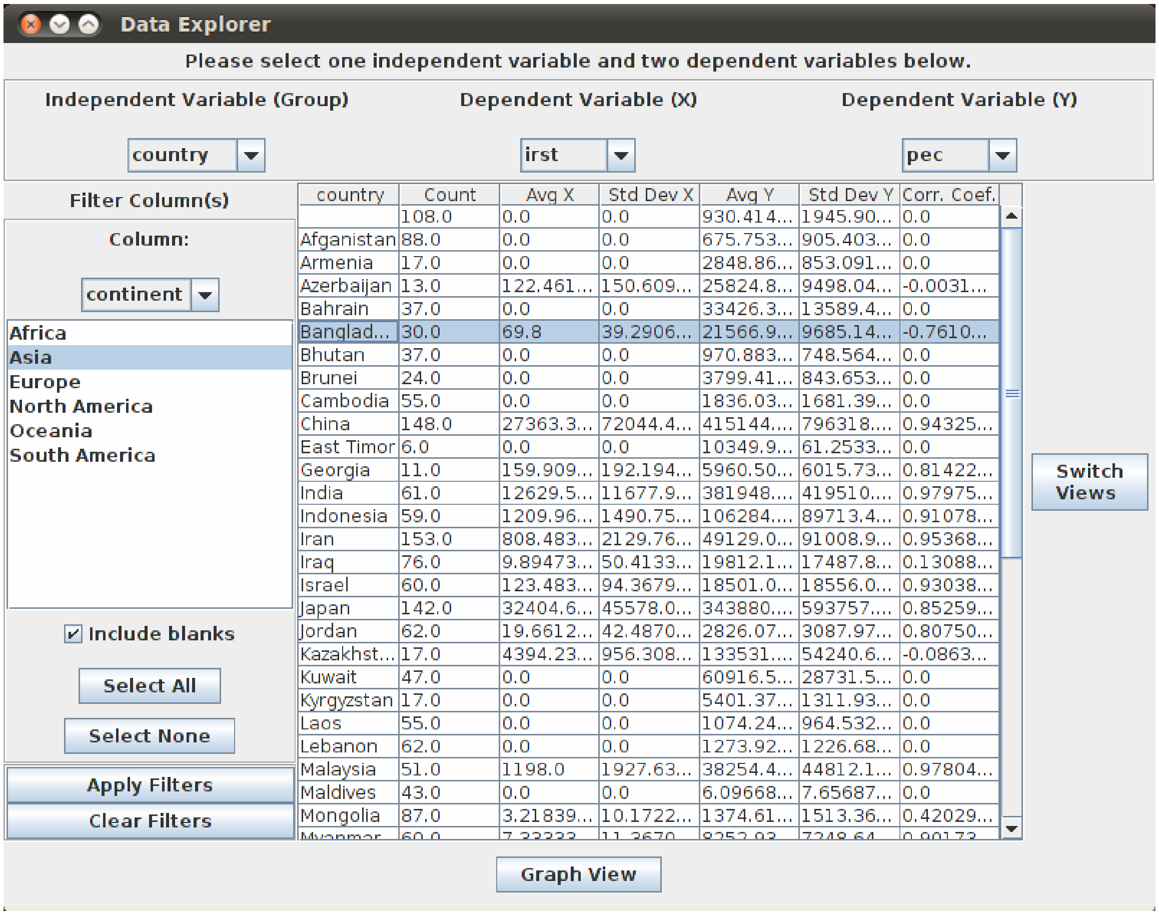
Column: continent

Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

Fig. 20. Primary Energy Consumption versus Iron and Steel Production by Country in Asia - Scatter Plot

China has the largest primary energy consumption and the largest iron and steel production as seen by the cyan series in Figure 20. To get a clearer view of the information in the chart, the data view will be helpful as seen in Figure 21.



Please select one independent variable and two dependent variables below.

Independent Variable (Group): country

Dependent Variable (X): irst

Dependent Variable (Y): pec

Filter Column(s)

Column: continent

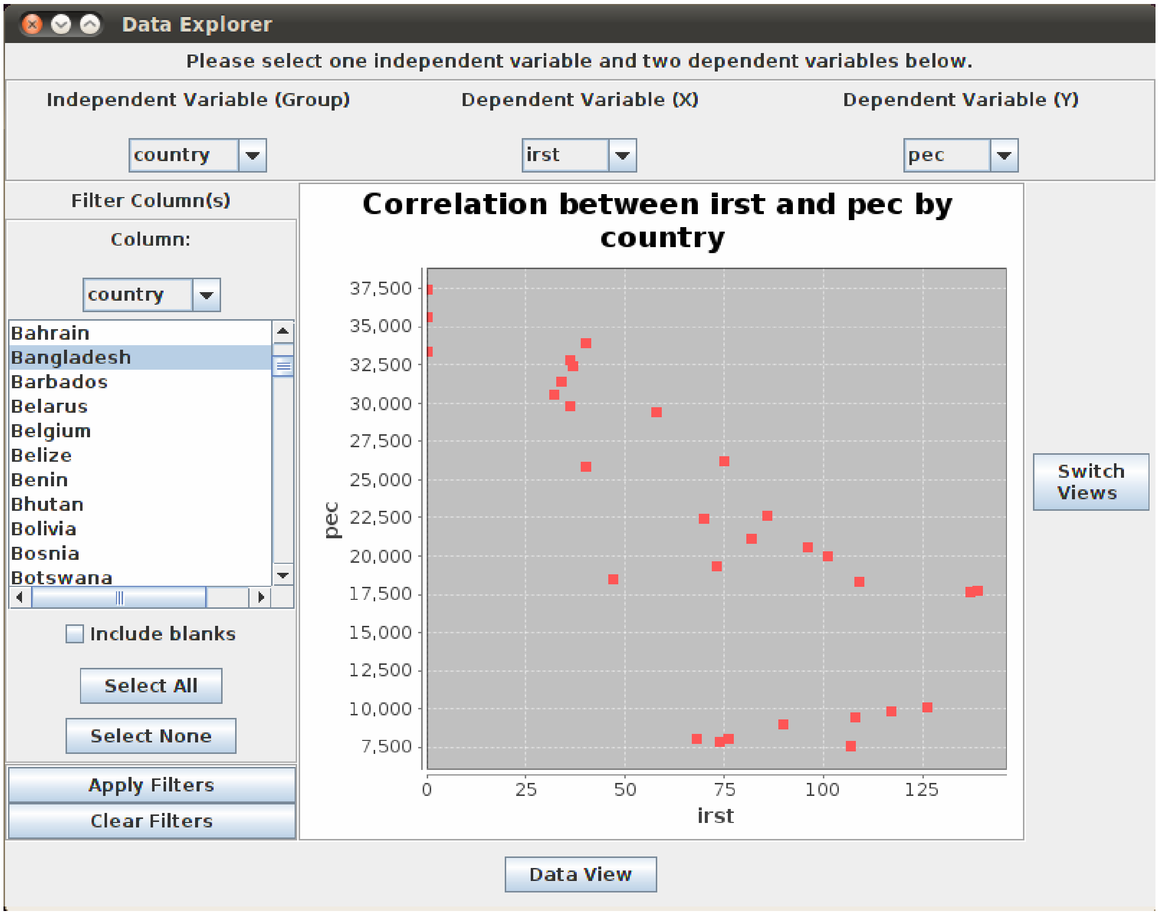
Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

country	Count	Avg X	Std Dev X	Avg Y	Std Dev Y	Corr. Coef.
	108.0	0.0	0.0	930.414...	1945.90...	0.0
Afganistan	88.0	0.0	0.0	675.753...	905.403...	0.0
Armenia	17.0	0.0	0.0	2848.86...	853.091...	0.0
Azerbaijan	13.0	122.461...	150.609...	25824.8...	9498.04...	-0.0031...
Bahrain	37.0	0.0	0.0	33426.3...	13589.4...	0.0
Banglad...	30.0	69.8	39.2906...	21566.9...	9685.14...	-0.7610...
Bhutan	37.0	0.0	0.0	970.883...	748.564...	0.0
Brunei	24.0	0.0	0.0	3799.41...	843.653...	0.0
Cambodia	55.0	0.0	0.0	1836.03...	1681.39...	0.0
China	148.0	27363.3...	72044.4...	415144....	796318....	0.94325...
East Timor	6.0	0.0	0.0	10349.9...	61.2533...	0.0
Georgia	11.0	159.909...	192.194...	5960.50...	6015.73...	0.81422...
India	61.0	12629.5...	11677.9...	381948....	419510....	0.97975...
Indonesia	59.0	1209.96...	1490.75...	106284....	89713.4...	0.91078...
Iran	153.0	808.483...	2129.76...	49129.0...	91008.9...	0.95368...
Iraq	76.0	9.89473...	50.4133...	19812.1...	17487.8...	0.13088...
Israel	60.0	123.483...	94.3679...	18501.0...	18556.0...	0.93038...
Japan	142.0	32404.6...	45578.0...	343880....	593757....	0.85259...
Jordan	62.0	19.6612...	42.4870...	2826.07...	3087.97...	0.80750...
Kazakhst...	17.0	4394.23...	956.308...	133531....	54240.6...	-0.0863...
Kuwait	47.0	0.0	0.0	60916.5...	28731.5...	0.0
Kyrgyzstan	17.0	0.0	0.0	5401.37...	1311.93...	0.0
Laos	55.0	0.0	0.0	1074.24...	964.532...	0.0
Lebanon	62.0	0.0	0.0	1273.92...	1226.68...	0.0
Malaysia	51.0	1198.0	1927.63...	38254.4...	44812.1...	0.97804...
Maldives	43.0	0.0	0.0	6.09668...	7.65687...	0.0
Mongolia	87.0	3.21839...	10.1722...	1374.61...	1513.36...	0.42029...
Myanmar	60.0	7.33333	11.3670	8252.93	7248.64	0.90173

Fig. 21. Primary Energy Consumption versus Iron and Steel Production by Country in Asia - Data View

Many countries have strong positive correlation coefficients between iron and steel production and primary energy consumption, while others have no apparent correlation. Some of the data is still incomplete resulting in no information for a variety of countries. Bangladesh is one of the few countries with a strong negative correlation coefficient, the highlighted row of data in Figure 21. Filtering the data to only Bangladesh and switching to the graph view can provide more information as seen in Figure 22.



Data Explorer

Please select one independent variable and two dependent variables below.

Independent Variable (Group): country

Dependent Variable (X): irst

Dependent Variable (Y): pec

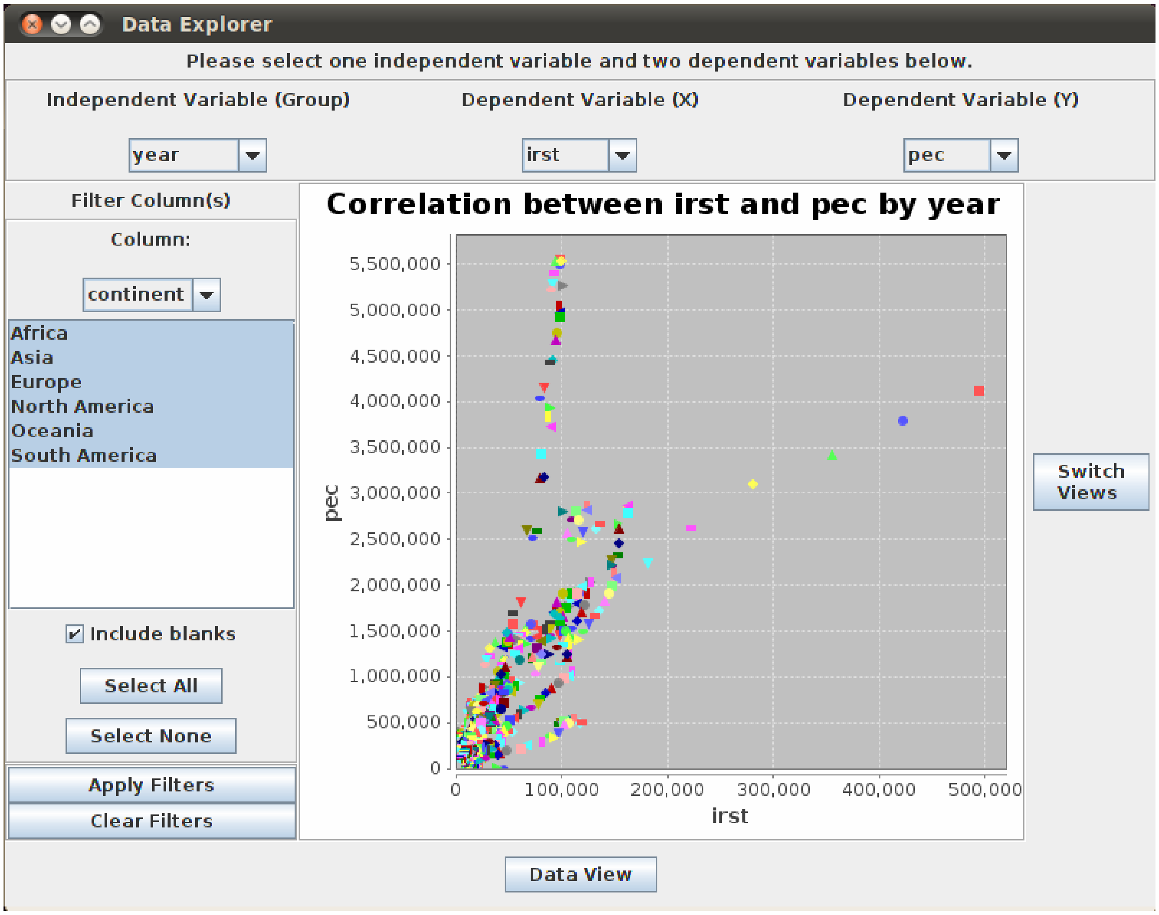
Filter Column(s)

Column: country

Fig. 22. Primary Energy Consumption versus Iron and Steel Production for Bangladesh – Scatter Plot

The negative coefficient means there is a negative linear relationship between iron and steel production and primary energy consumption. The closer the value is to 1 or -1 the stronger the relationship. The correlation coefficient indeed matches the scatter plot view, though the scatter plot of only Bangladesh in Figure 22 is much easier to read than the scatter plot of Asia as a whole in Figure 20.

IRST and PEC Correlation by Year



Please select one independent variable and two dependent variables below.

Independent Variable (Group): year

Dependent Variable (X): irst

Dependent Variable (Y): pec

Filter Column(s)

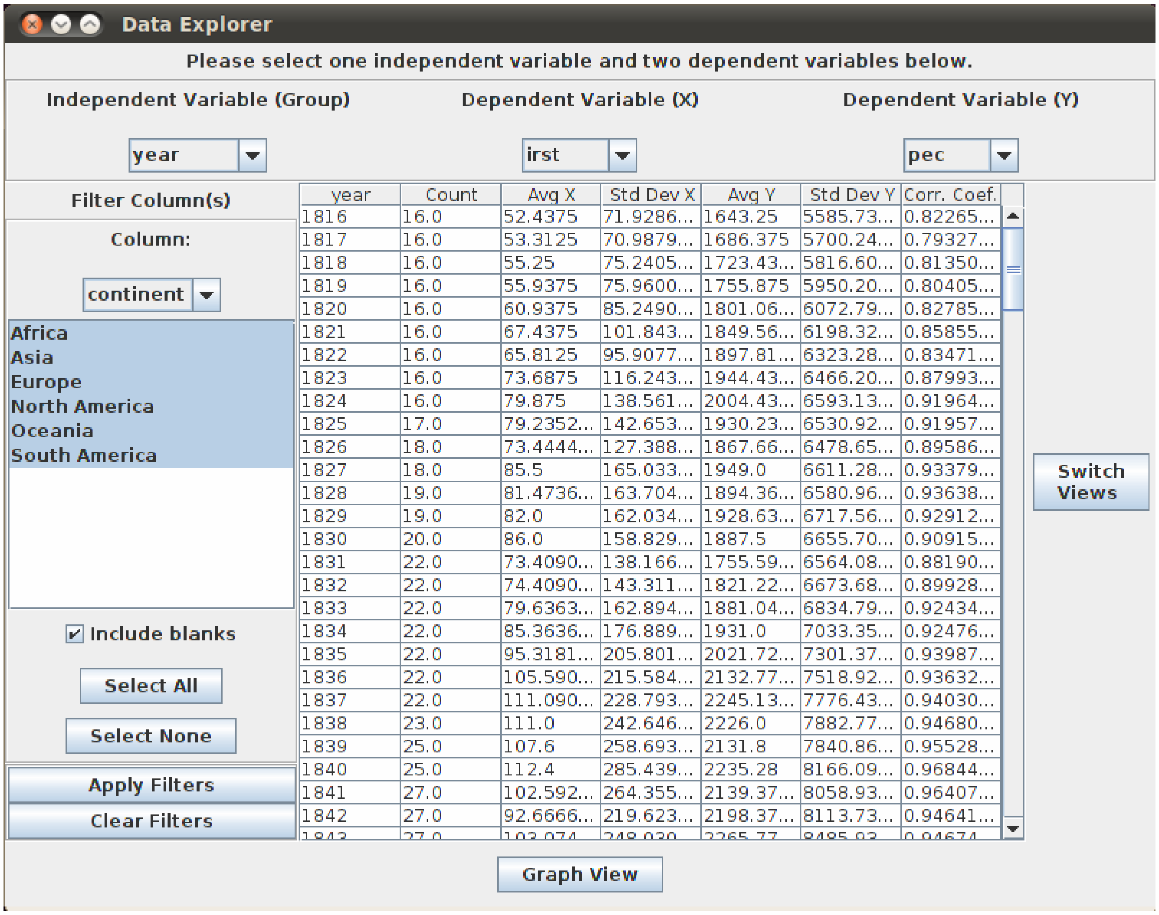
Column: continent

Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

Fig. 23. Primary Energy Consumption versus Iron and Steel Production by Year – Scatter Plot

The correlation between iron and steel production and primary energy consumption grouped by year is not clear from the scatter plot in Figure 23. The plot does not provide discernible information. The data view will be more appropriate, as shown in Figures 24 and 25.



Data Explorer

Please select one independent variable and two dependent variables below.

Independent Variable (Group): year

Dependent Variable (X): irst

Dependent Variable (Y): pec

Filter Column(s)

Column: continent

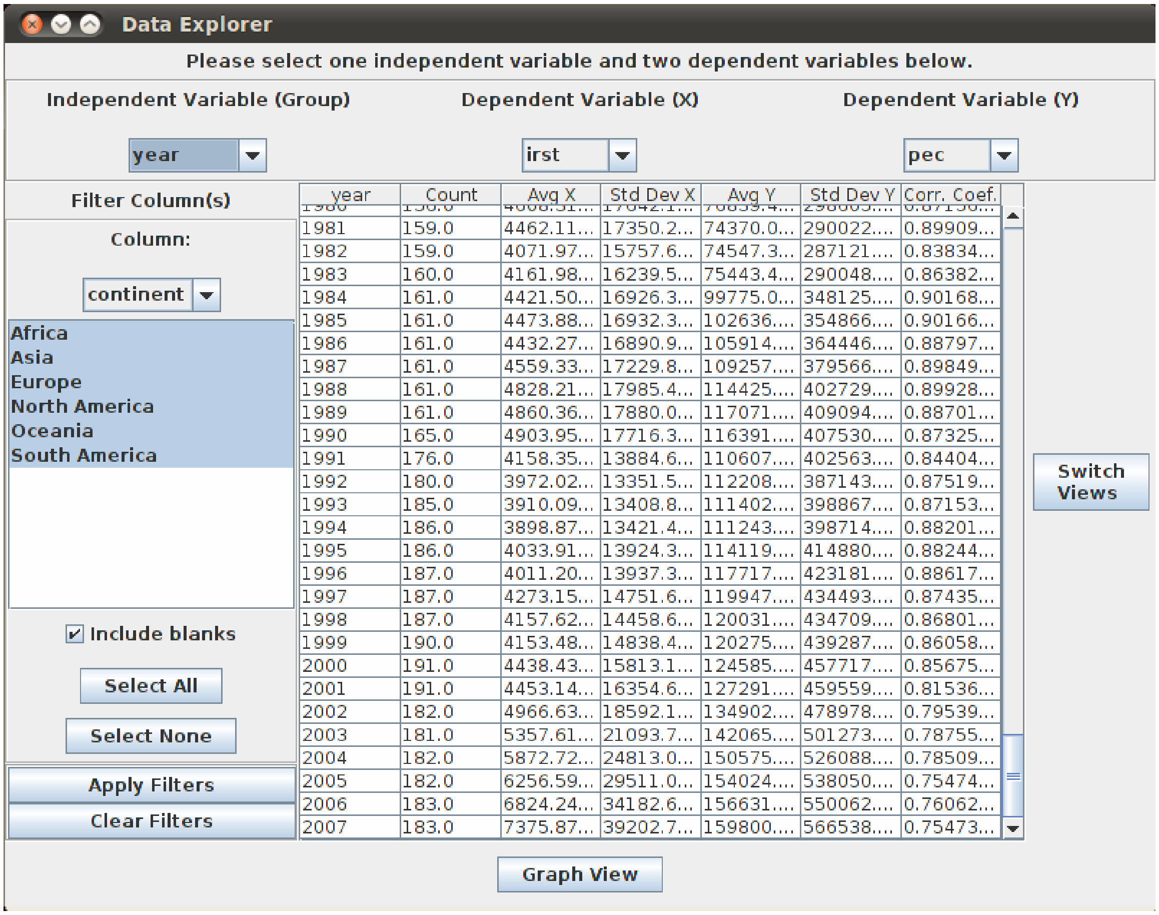
Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

year	Count	Avg X	Std Dev X	Avg Y	Std Dev Y	Corr. Coef.
1816	16.0	52.4375	71.9286...	1643.25	5585.73...	0.82265...
1817	16.0	53.3125	70.9879...	1686.375	5700.24...	0.79327...
1818	16.0	55.25	75.2405...	1723.43...	5816.60...	0.81350...
1819	16.0	55.9375	75.9600...	1755.875	5950.20...	0.80405...
1820	16.0	60.9375	85.2490...	1801.06...	6072.79...	0.82785...
1821	16.0	67.4375	101.843...	1849.56...	6198.32...	0.85855...
1822	16.0	65.8125	95.9077...	1897.81...	6323.28...	0.83471...
1823	16.0	73.6875	116.243...	1944.43...	6466.20...	0.87993...
1824	16.0	79.875	138.561...	2004.43...	6593.13...	0.91964...
1825	17.0	79.2352...	142.653...	1930.23...	6530.92...	0.91957...
1826	18.0	73.4444...	127.388...	1867.66...	6478.65...	0.89586...
1827	18.0	85.5	165.033...	1949.0	6611.28...	0.93379...
1828	19.0	81.4736...	163.704...	1894.36...	6580.96...	0.93638...
1829	19.0	82.0	162.034...	1928.63...	6717.56...	0.92912...
1830	20.0	86.0	158.829...	1887.5	6655.70...	0.90915...
1831	22.0	73.4090...	138.166...	1755.59...	6564.08...	0.88190...
1832	22.0	74.4090...	143.311...	1821.22...	6673.68...	0.89928...
1833	22.0	79.6363...	162.894...	1881.04...	6834.79...	0.92434...
1834	22.0	85.3636...	176.889...	1931.0	7033.35...	0.92476...
1835	22.0	95.3181...	205.801...	2021.72...	7301.37...	0.93987...
1836	22.0	105.590...	215.584...	2132.77...	7518.92...	0.93632...
1837	22.0	111.090...	228.793...	2245.13...	7776.43...	0.94030...
1838	23.0	111.0	242.646...	2226.0	7882.77...	0.94680...
1839	25.0	107.6	258.693...	2131.8	7840.86...	0.95528...
1840	25.0	112.4	285.439...	2235.28	8166.09...	0.96844...
1841	27.0	102.592...	264.355...	2139.37...	8058.93...	0.96407...
1842	27.0	92.6666...	219.623...	2198.37...	8113.73...	0.94641...
1843	27.0	103.074	248.030	2265.77	8485.93	0.94674

Fig. 24. Primary Energy Consumption versus Iron and Steel Production by Year 1816-1842 – Data View

Starting in 1816, the correlation between the two variables, iron and steel production and primary energy consumption, increases over time shown by the Corr. Coef. column in Figure 24. This makes sense that the more iron and steel produced, the more energy is consumed and vice versa.



Please select one independent variable and two dependent variables below.

Independent Variable (Group): year

Dependent Variable (X): irst

Dependent Variable (Y): pec

Filter Column(s)

Column: continent

Africa
Asia
Europe
North America
Oceania
South America

☑ Include blanks

year	Count	Avg X	Std Dev X	Avg Y	Std Dev Y	Corr. Coef.
1981	159.0	4462.11...	17350.2...	74370.0...	290022....	0.89909...
1982	159.0	4071.97...	15757.6...	74547.3...	287121....	0.83834...
1983	160.0	4161.98...	16239.5...	75443.4...	290048....	0.86382...
1984	161.0	4421.50...	16926.3...	99775.0...	348125....	0.90168...
1985	161.0	4473.88...	16932.3...	102636....	354866....	0.90166...
1986	161.0	4432.27...	16890.9...	105914....	364446....	0.88797...
1987	161.0	4559.33...	17229.8...	109257....	379566....	0.89849...
1988	161.0	4828.21...	17985.4...	114425....	402729....	0.89928...
1989	161.0	4860.36...	17880.0...	117071....	409094....	0.88701...
1990	165.0	4903.95...	17716.3...	116391....	407530....	0.87325...
1991	176.0	4158.35...	13884.6...	110607....	402563....	0.84404...
1992	180.0	3972.02...	13351.5...	112208....	387143....	0.87519...
1993	185.0	3910.09...	13408.8...	111402....	398867....	0.87153...
1994	186.0	3898.87...	13421.4...	111243....	398714....	0.88201...
1995	186.0	4033.91...	13924.3...	114119....	414880....	0.88244...
1996	187.0	4011.20...	13937.3...	117717....	423181....	0.88617...
1997	187.0	4273.15...	14751.6...	119947....	434493....	0.87435...
1998	187.0	4157.62...	14458.6...	120031....	434709....	0.86801...
1999	190.0	4153.48...	14838.4...	120275....	439287....	0.86058...
2000	191.0	4438.43...	15813.1...	124585....	457717....	0.85675...
2001	191.0	4453.14...	16354.6...	127291....	459559....	0.81536...
2002	182.0	4966.63...	18592.1...	134902....	478978....	0.79539...
2003	181.0	5357.61...	21093.7...	142065....	501273....	0.78755...
2004	182.0	5872.72...	24813.0...	150575....	526088....	0.78509...
2005	182.0	6256.59...	29511.0...	154024....	538050....	0.75474...
2006	183.0	6824.24...	34182.6...	156631....	550062....	0.76062...
2007	183.0	7375.87...	39202.7...	159800....	566538....	0.75473...

Fig. 25. Primary Energy Consumption versus Iron and Steel Production by Year 1981-2007 – Data View

However, the more recent data indicates that the correlation between the two variables is now decreasing in Figure 25. This is an unexpected but interesting trend. Primary energy consumption is now used for other purposes, rather than only iron and steel production.

Case 2: IT Service Support Management for Company X

For any organization with a division providing IT services, the support of those services can often be a time consuming and expensive. A large networking company, Company X, has integrated various monitoring systems on their IT support process, to ensure service outages of any kind are resolved quickly and efficiently.

A snapshot of data for one fiscal week was taken, containing over 170 columns and 160,000 rows of data. The snapshot was scrubbed of columns with identifying company information and redundant columns from the database of information. A 15,000 row subset of the resulting data was used as a sample. Each row represents the information about one incident, or service outage.

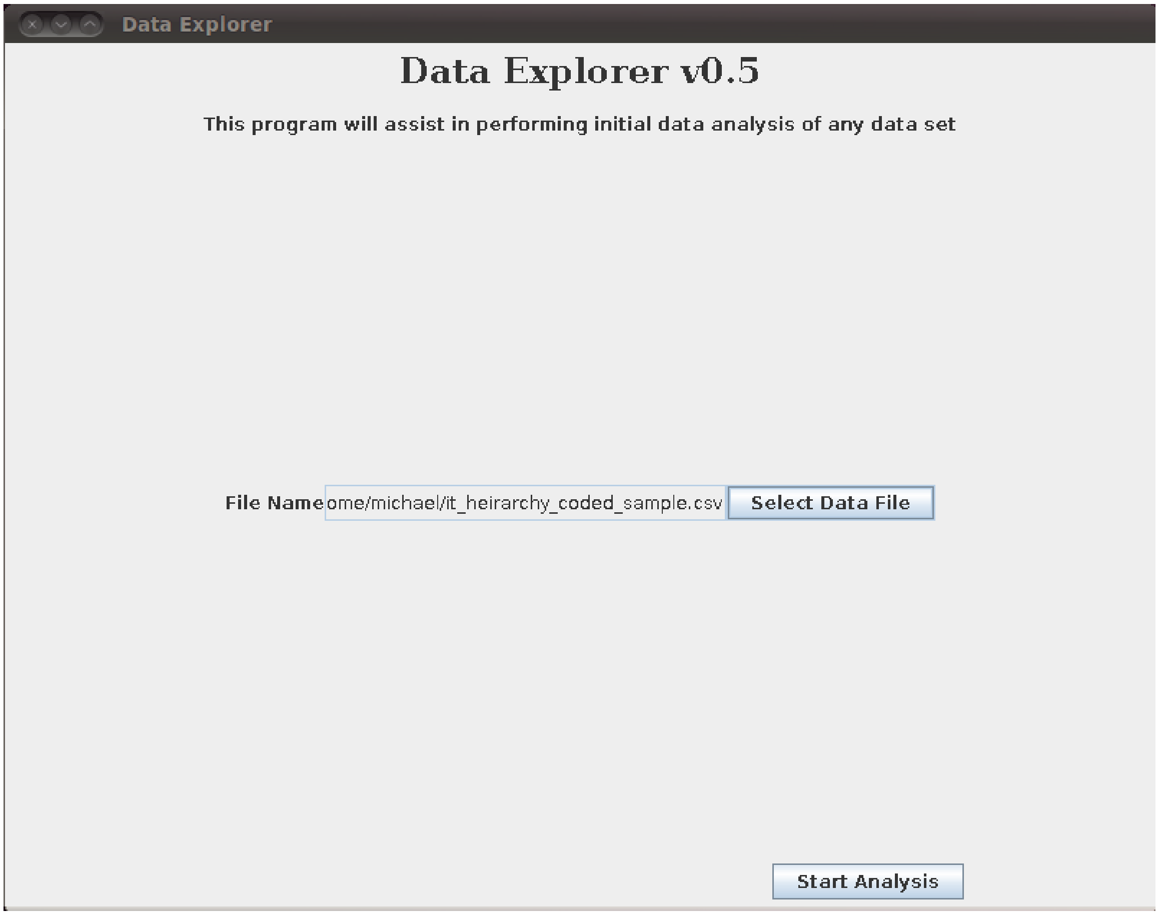
This data set was selected based on its applicability to the service industry model. Many organizations are moving towards selling products as services rather than single purchases. The large networking company operates on this model and was willing to provide data, so long as all identifying information from the data was removed.

The following information, graphs, and data were obtained using the prototype implementation of the methodology. Figures 26 through 45 are screenshots taken directly from the prototype implementation and can be used as a guide to replicate the analysis.

The largest advantage the methodology provides for this data set is the ability to quickly change the view of the data within a hierarchy. In a large organization, the information desired by an executive will be very different than the information desired by an analyst. The methodology and prototype implementation allow the user to change his or her perspective, gaining insight into other areas of the organization.

Manually creating just one of the charts in Figures 26 through 45 and manipulating the data in spreadsheet software would take close to an hour on such a large data set. Using the prototype implementation of the methodology, the change of perspective in any data set requires at most three mouse clicks, one to change filter the current level in the hierarchy to a specific value, one to apply the filter, and one to change the independent variable to the next lowest level in the hierarchy. These three clicks and the automated execution of the tasks take seconds for this data set. This significantly reduces the time spent by the user to generate the graphs and allows them to shift focus to gain additional perspective. The user can ask and answer more questions, improving the quality of the analysis.

Phase 1: Data Acquisition



Data Explorer v0.5

This program will assist in performing initial data analysis of any data set

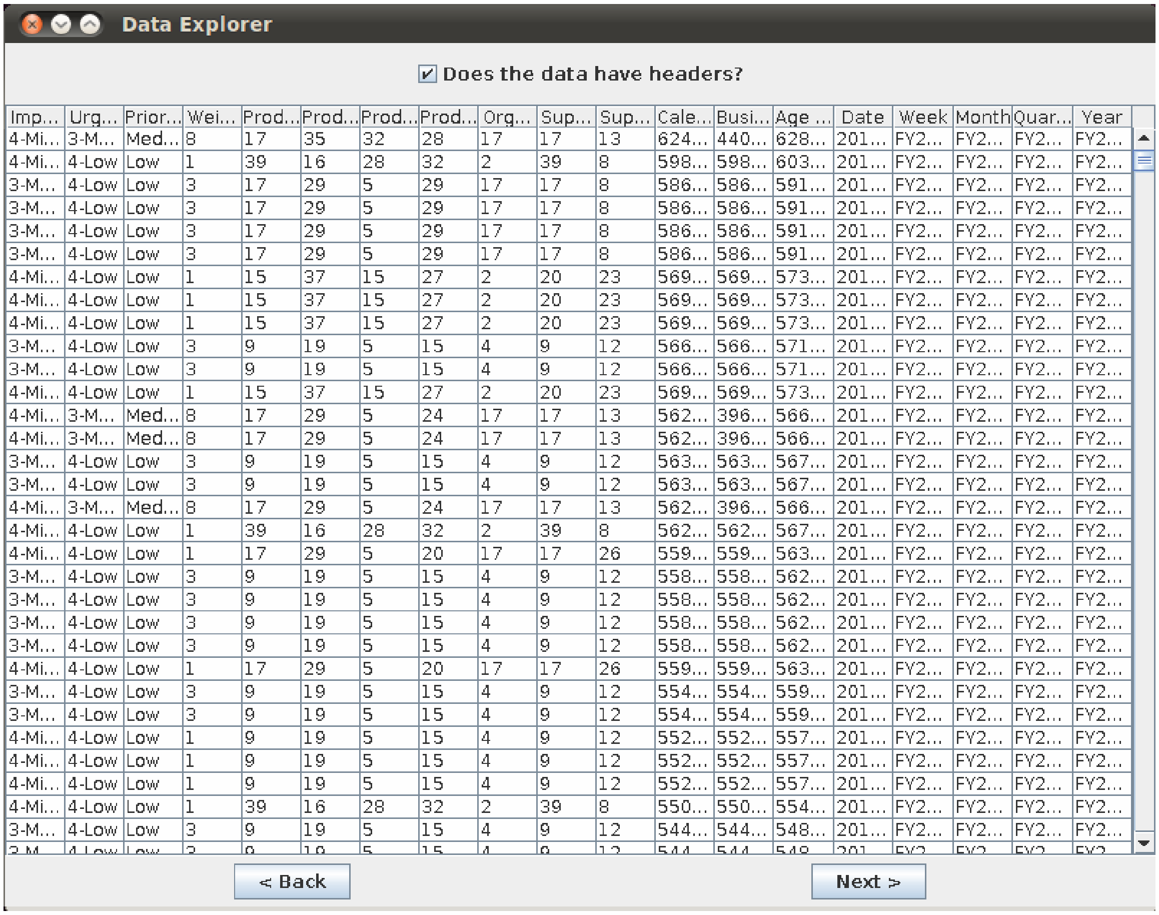
File Name ome/michael/it_heirarchy_coded_sample.csv

Fig. 26. IT Data File Selected

Again, similar to Figure 5, Figure 26 shows that once a file has been selected the analysis can begin. For comparison, see Figure 3 to view the disabled version of the button.

Phase 2: Data Profiling

Internal Data File Structure



☑ Does the data have headers?

Imp...	Urg...	Prior...	Wei...	Prod...	Prod...	Prod...	Prod...	Org...	Sup...	Sup...	Cale...	Busi...	Age ...	Date	Week	Month	Quar...	Year
4-Mi...	3-M...	Med...	8	17	35	32	28	17	17	13	624...	440...	628...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	39	16	28	32	2	39	8	598...	598...	603...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	566...	566...	571...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	566...	566...	571...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	3-M...	Med...	8	17	29	5	24	17	17	13	562...	396...	566...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	3-M...	Med...	8	17	29	5	24	17	17	13	562...	396...	566...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	563...	563...	567...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	563...	563...	567...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	3-M...	Med...	8	17	29	5	24	17	17	13	562...	396...	566...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	39	16	28	32	2	39	8	562...	562...	567...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	17	29	5	20	17	17	26	559...	559...	563...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	17	29	5	20	17	17	26	559...	559...	563...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	554...	554...	559...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	554...	554...	559...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	9	19	5	15	4	9	12	552...	552...	557...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	9	19	5	15	4	9	12	552...	552...	557...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	9	19	5	15	4	9	12	552...	552...	557...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	39	16	28	32	2	39	8	550...	550...	554...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	544...	544...	548...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	544...	544...	548...	201...	FY2...	FY2...	FY2...	FY2...

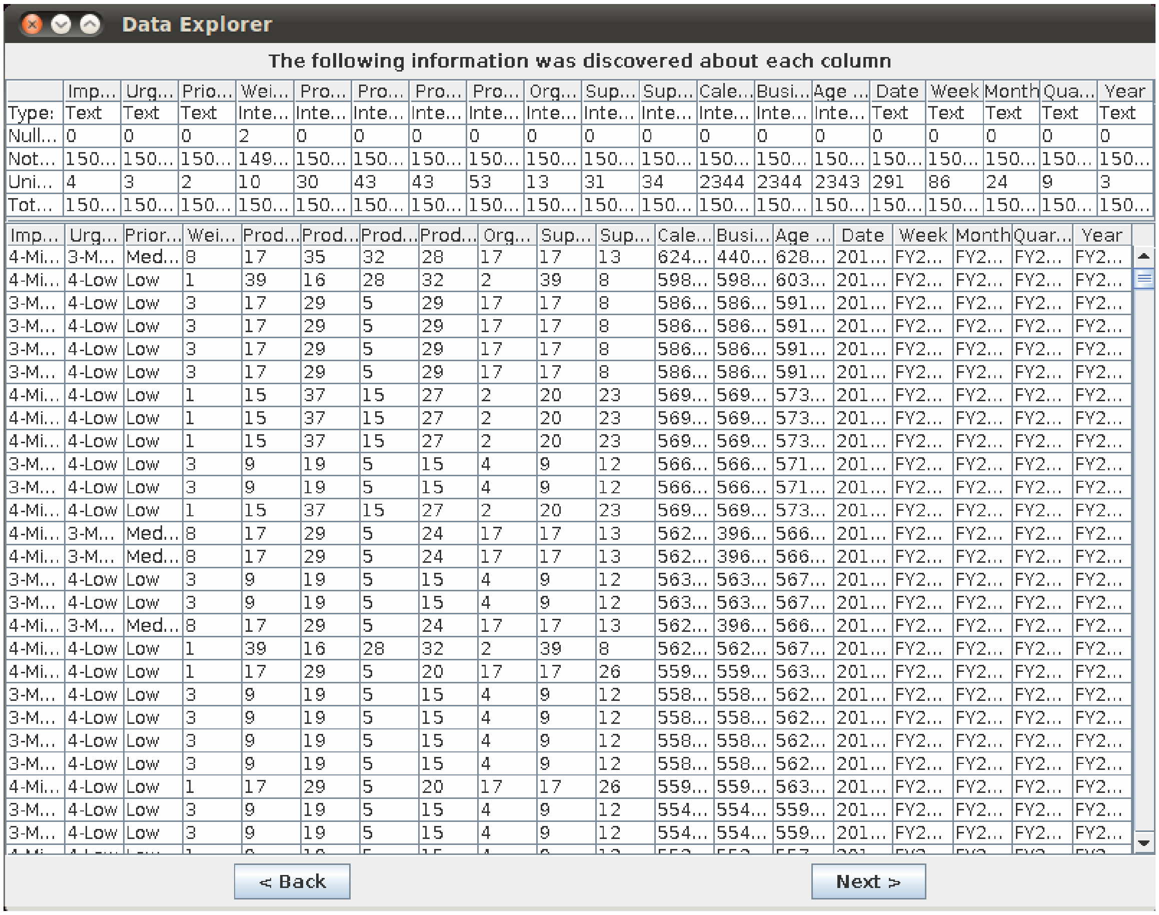
Fig. 27. IT Headers Verified

The data file has headers with some descriptive information. The product and organization columns have been coded to preserve the hierarchical structure while removing identifying information.

The headers in Figure 27 correspond to the following information:

- Impact – the extent the outage is affecting the population of users
- Urgency – the need for the outage to be fixed quickly
- Priority – a combination of the impact and urgency
- Weight – the weight of the specific priority
- Product T1 – the first tier of product categories
- Product T2 – the second tier of product categories
- Product T3 – the third tier of product categories
- Product – the specific product related to the incident
- Organization – the larger organization handling the incident
- Support Org. – the support organization handling the incident
- Support Group – the support team handling the incident
- Calendar Resolution Time (sec) – the time in seconds to completely resolve the incident
- Business Resolution Time (sec) – the time in seconds to eliminate any business impacts an incident may be causing
- Age (sec) – the time in seconds since the incident was opened
- Date – the date an incident was logged
- Week – the week an incident was logged
- Month – the month an incident was logged
- Quarter – the quarter an incident was logged
- Year – the year an incident was logged

Type Checking, Missing Values, and Individual Column Cardinality



The following information was discovered about each column

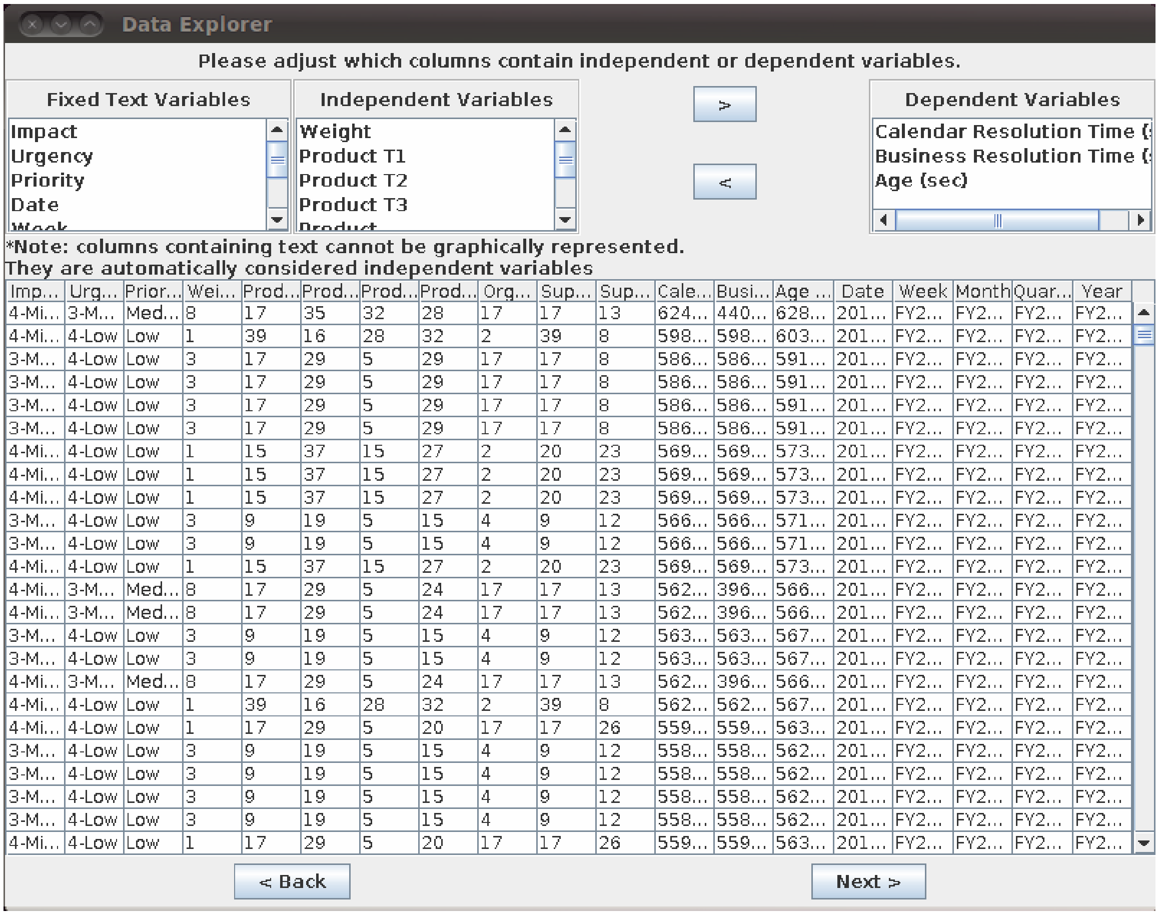
	Imp…	Urg…	Prio…	Wei…	Pro…	Pro…	Pro…	Pro…	Org…	Sup…	Sup…	Cale…	Busi…	Age …	Date	Week	Month	Qua…	Year
Type:	Text	Text	Text	Inte…	Inte…	Inte…	Inte…	Inte…	Inte…	Inte…	Inte…	Inte…	Inte…	Inte…	Text	Text	Text	Text	Text
Null…	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Not…	150…	150…	150…	149…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…
Uni…	4	3	2	10	30	43	43	53	13	31	34	2344	2344	2343	291	86	24	9	3
Tot…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…	150…

Imp…	Urg…	Prior…	Wei…	Prod…	Prod…	Prod…	Prod…	Org…	Sup…	Sup…	Cale…	Busi…	Age …	Date	Week	Month	Quar…	Year
4-Mi…	3-M…	Med…	8	17	35	32	28	17	17	13	624…	440…	628…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	39	16	28	32	2	39	8	598…	598…	603…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	566…	566…	571…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	566…	566…	571…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	3-M…	Med…	8	17	29	5	24	17	17	13	562…	396…	566…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	3-M…	Med…	8	17	29	5	24	17	17	13	562…	396…	566…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	563…	563…	567…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	563…	563…	567…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	3-M…	Med…	8	17	29	5	24	17	17	13	562…	396…	566…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	39	16	28	32	2	39	8	562…	562…	567…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	17	29	5	20	17	17	26	559…	559…	563…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	17	29	5	20	17	17	26	559…	559…	563…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	554…	554…	559…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	554…	554…	559…	201…	FY2…	FY2…	FY2…	FY2…

Fig. 28. IT Column Summaries

Figure 28 shows that despite the very large number of rows in the data set, only three columns have more than 300 unique values. This suggests that the data could be normalized into data base tables to compress the information. The Impact, Urgency, Priority, Weight, Product T1, Product T2, Product T3, Product, Date, Week, Quarter, and Year could be stored in separate tables. A key could be used to tie the specific values to the separate table, reducing the size of the data. The relationships between the columns would be needed to perform this normalization, shown in Figure 30.

Column Dependence



Please adjust which columns contain independent or dependent variables.

Fixed Text Variables
Impact
Urgency
Priority
Date
Week

Independent Variables
Weight
Product T1
Product T2
Product T3
Product

Dependent Variables
Calendar Resolution Time (
Business Resolution Time (
Age (sec)

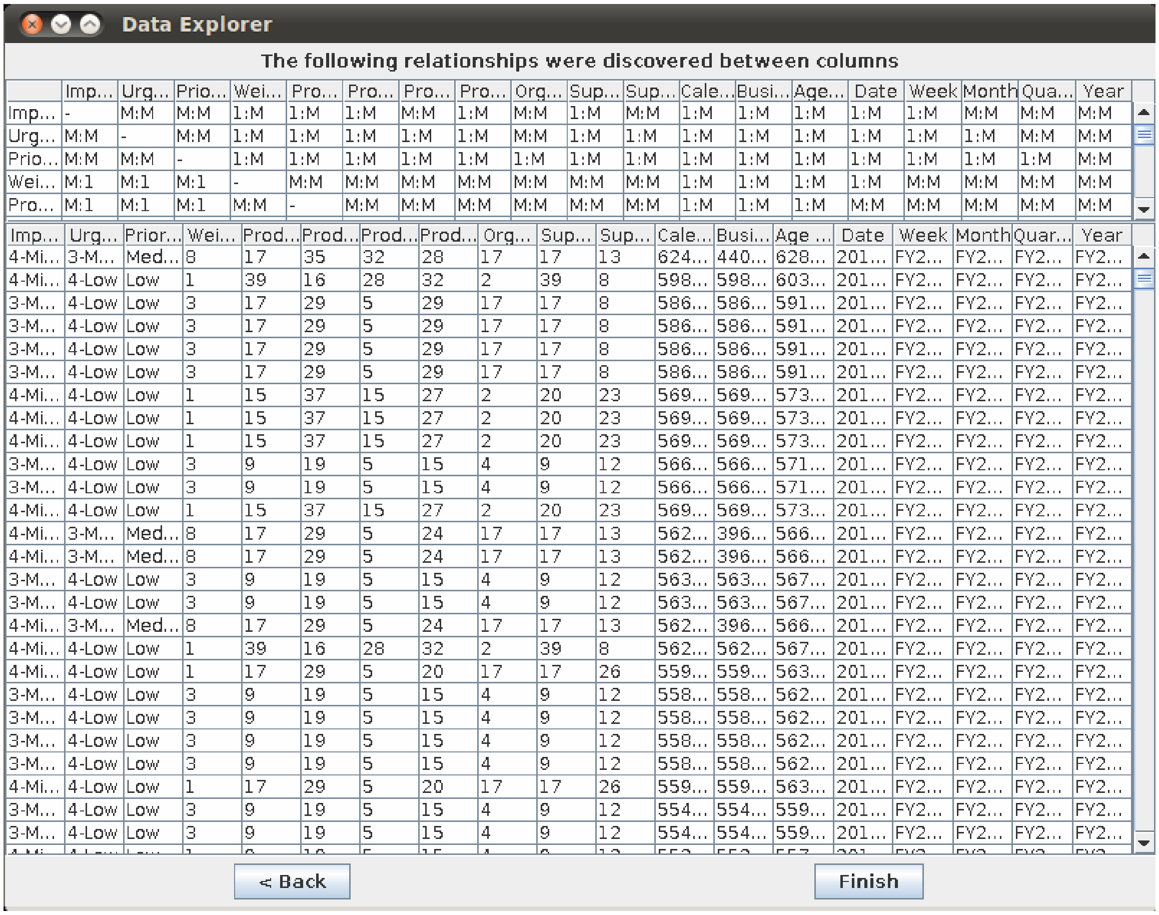
*Note: columns containing text cannot be graphically represented.
They are automatically considered independent variables

Imp…	Urg…	Prior…	Wei…	Prod…	Prod…	Prod…	Prod…	Org…	Sup…	Sup…	Cale…	Busi…	Age …	Date	Week	Month	Quar…	Year
4-Mi…	3-M…	Med…	8	17	35	32	28	17	17	13	624…	440…	628…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	39	16	28	32	2	39	8	598…	598…	603…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	17	29	5	29	17	17	8	586…	586…	591…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	566…	566…	571…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	566…	566…	571…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	15	37	15	27	2	20	23	569…	569…	573…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	3-M…	Med…	8	17	29	5	24	17	17	13	562…	396…	566…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	3-M…	Med…	8	17	29	5	24	17	17	13	562…	396…	566…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	563…	563…	567…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	563…	563…	567…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	3-M…	Med…	8	17	29	5	24	17	17	13	562…	396…	566…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	39	16	28	32	2	39	8	562…	562…	567…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	17	29	5	20	17	17	26	559…	559…	563…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
3-M…	4-Low	Low	3	9	19	5	15	4	9	12	558…	558…	562…	201…	FY2…	FY2…	FY2…	FY2…
4-Mi…	4-Low	Low	1	17	29	5	20	17	17	26	559…	559…	563…	201…	FY2…	FY2…	FY2…	FY2…

Fig. 29. IT Independent & Dependent Variable Selection

The coded product tiers and organizations are all represented as numbers. They can and should be used as independent variables to investigate product and organization specific information. Figure 29 shows the independent and dependent variables after adjustment through the use of the toggle buttons.

Column Relationships



The following relationships were discovered between columns

	Imp...	Urg...	Prio...	Wei...	Pro...	Pro...	Pro...	Pro...	Org...	Sup...	Sup...	Cale...	Busi...	Age...	Date	Week	Month	Qua...	Year
Imp...	-	M:M	M:M	1:M	1:M	1:M	M:M	1:M	M:M	1:M	M:M	1:M	1:M	1:M	1:M	1:M	M:M	M:M	M:M
Urg...	M:M	-	M:M	1:M	1:M	1:M	1:M	1:M	M:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	M:M	M:M
Prio...	M:M	M:M	-	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	1:M	M:M
Wei...	M:1	M:1	M:1	-	M:M	M:M	M:M	M:M	M:M	M:M	M:M	1:M	1:M	1:M	1:M	M:M	M:M	M:M	M:M
Pro...	M:1	M:1	M:1	M:M	-	M:M	M:M	M:M	M:M	M:M	M:M	1:M	1:M	1:M	M:M	M:M	M:M	M:M	M:M

Imp...	Urg...	Prior...	Wei...	Prod...	Prod...	Prod...	Prod...	Org...	Sup...	Sup...	Cale...	Busi...	Age ...	Date	Week	Month	Quar...	Year
4-Mi...	3-M...	Med...	8	17	35	32	28	17	17	13	624...	440...	628...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	39	16	28	32	2	39	8	598...	598...	603...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	17	29	5	29	17	17	8	586...	586...	591...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	566...	566...	571...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	566...	566...	571...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	15	37	15	27	2	20	23	569...	569...	573...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	3-M...	Med...	8	17	29	5	24	17	17	13	562...	396...	566...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	3-M...	Med...	8	17	29	5	24	17	17	13	562...	396...	566...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	563...	563...	567...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	563...	563...	567...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	3-M...	Med...	8	17	29	5	24	17	17	13	562...	396...	566...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	39	16	28	32	2	39	8	562...	562...	567...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	17	29	5	20	17	17	26	559...	559...	563...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	558...	558...	562...	201...	FY2...	FY2...	FY2...	FY2...
4-Mi...	4-Low	Low	1	17	29	5	20	17	17	26	559...	559...	563...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	554...	554...	559...	201...	FY2...	FY2...	FY2...	FY2...
3-M...	4-Low	Low	3	9	19	5	15	4	9	12	554...	554...	559...	201...	FY2...	FY2...	FY2...	FY2...

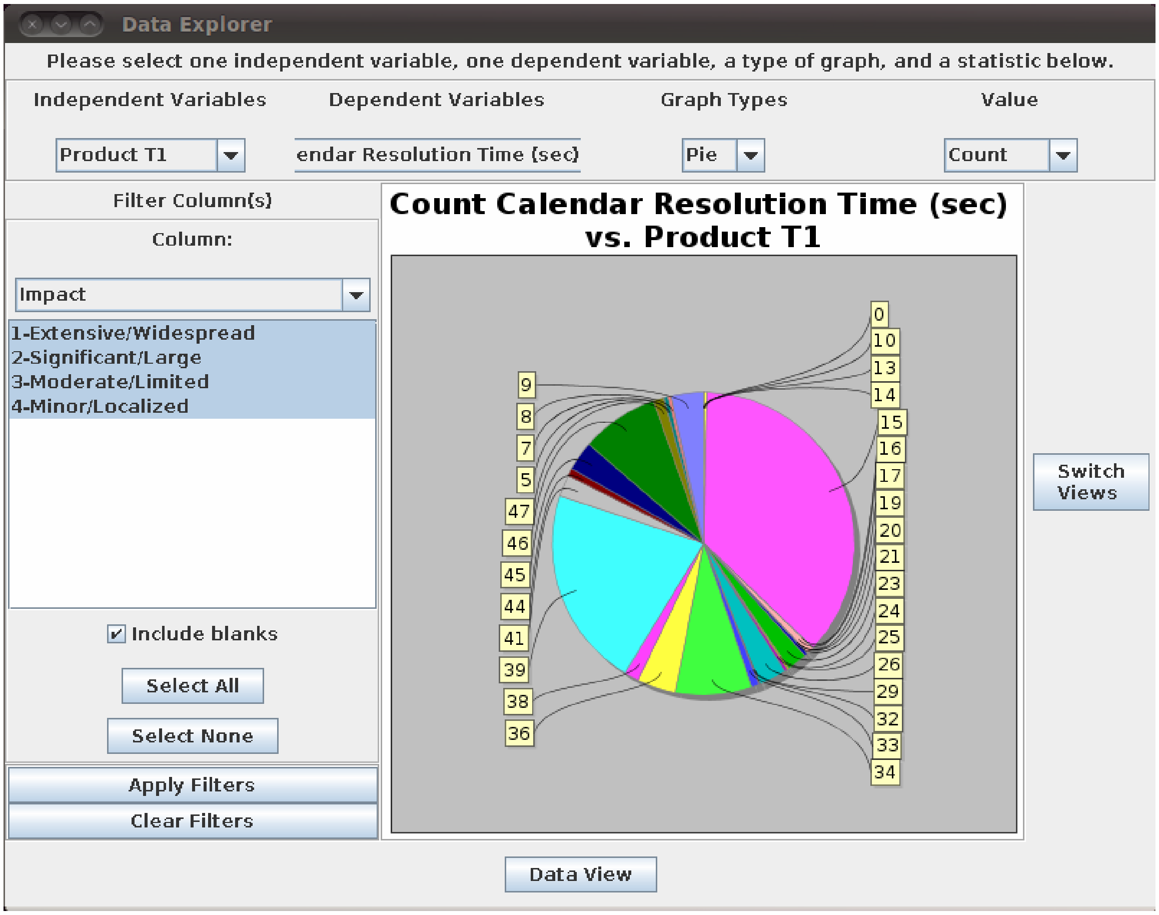
Fig. 30. IT Column Relationships

Each Impact, Urgency, or Priority identifies a single Weight, as seen by the 1:M representing a one to many relationship in Figure 30. There is also a hierarchy between the Date, Week, Month, Quarter, and Year columns. The product tiers and organizations have implied hierarchies.

Phase 3: Data Summaries & Visualizations

Incident Counts by Product

The number of incidents for a specific product is particularly relevant to managing the workload of support teams and identifying which products need additional review for quality. Figures 31 through 34 show the progression of pie charts with more specific levels of filters, moving from the Product T1, to the Product T2, to the Product T3, to the Products. The graphs filter based on the previous category which was the largest proportion of the total number of incidents. For the three sets of product tiers, it becomes clear there are a few products which are causing a large number of incidents. Product T3 27 is the largest contributor of incidents in the sample data. The final view, Figure 34, shows the Products which are contributing the largest number of incidents within Product T3 27.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Product T1

Dependent Variables: endar Resolution Time (sec)

Graph Types: Pie

Value: Count

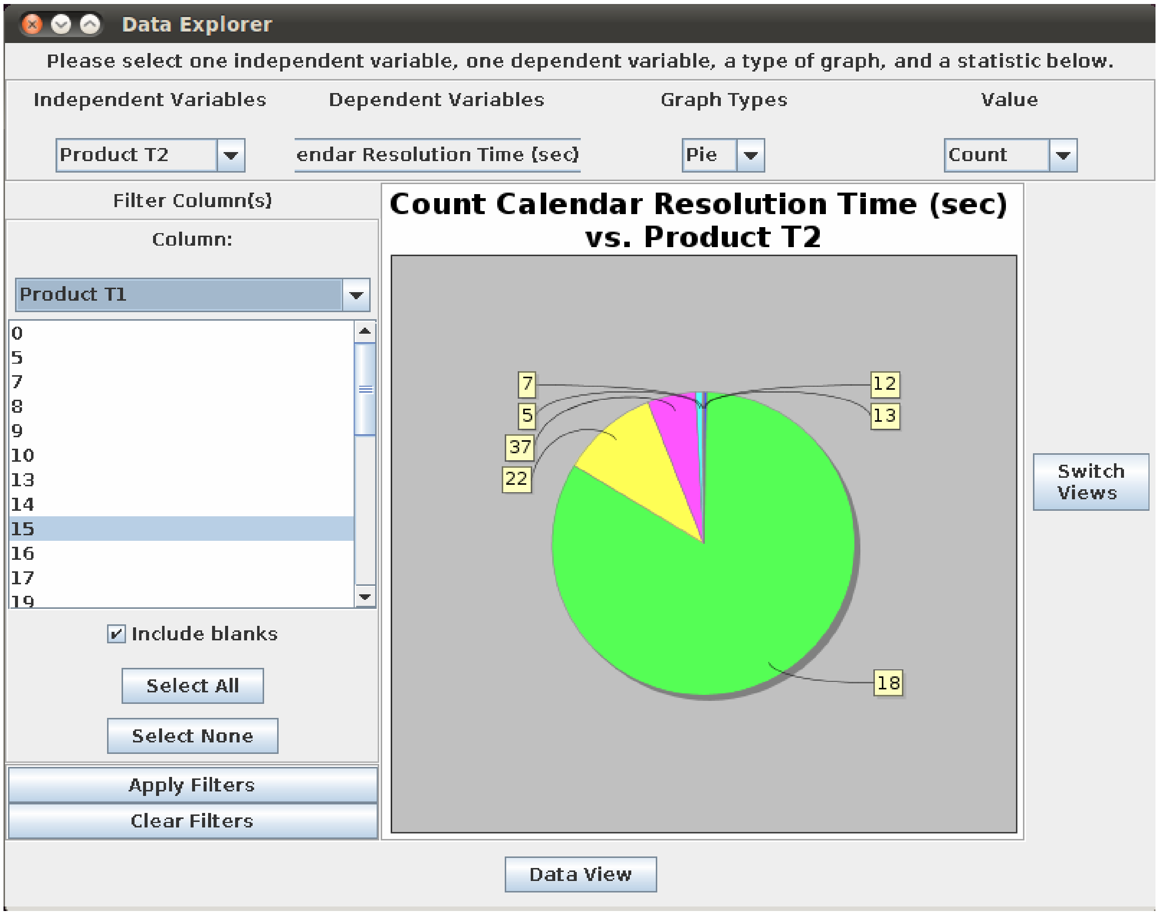
Filter Column(s)

Column: Impact

1-Extensive/Widespread
2-Significant/Large
3-Moderate/Limited
4-Minor/Localized

☑ Include blanks

Fig. 31. Count of Incidents by Product T1 - Pie Chart



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Product T2

Dependent Variables: endar Resolution Time (sec)

Graph Types: Pie

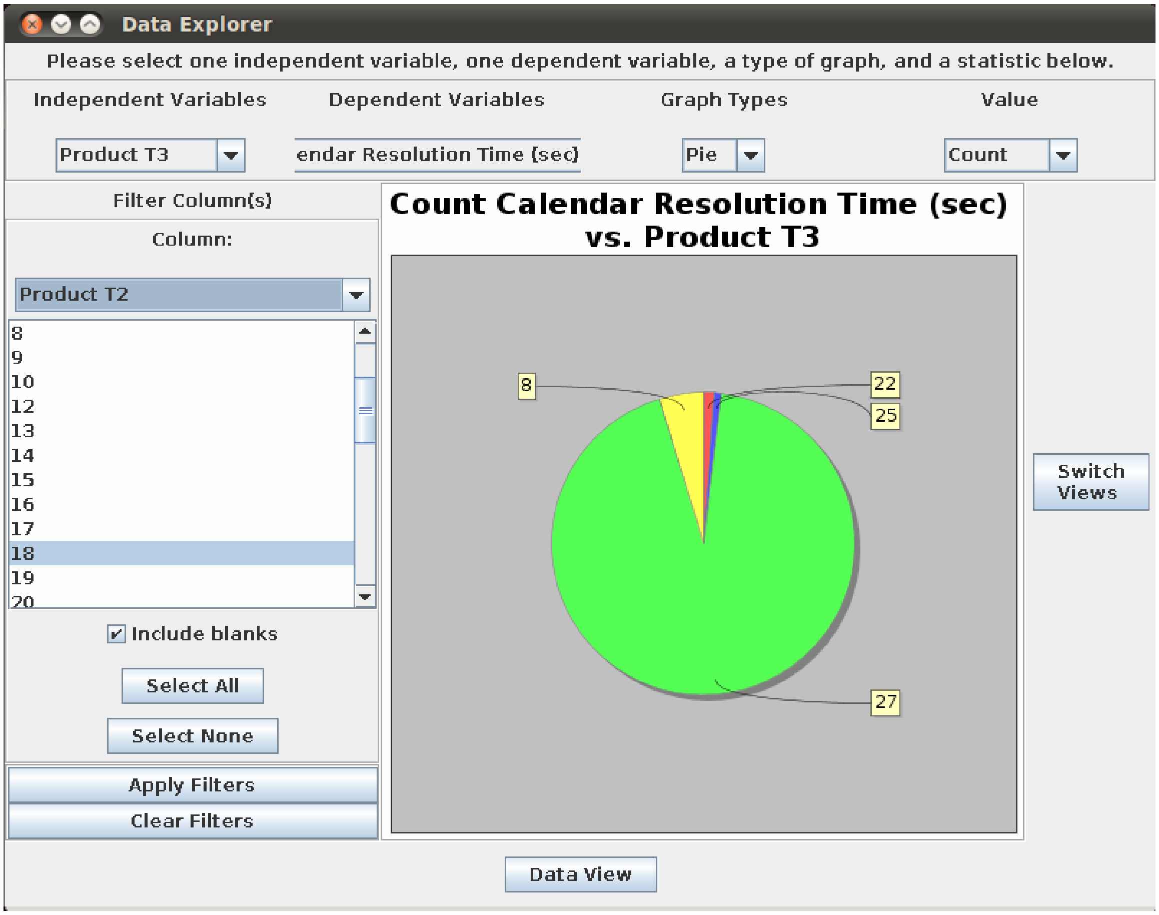
Value: Count

Filter Column(s)

Column: Product T1

☑ Include blanks

Fig. 32. Count of Incidents by Product T2 - Pie Chart



Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

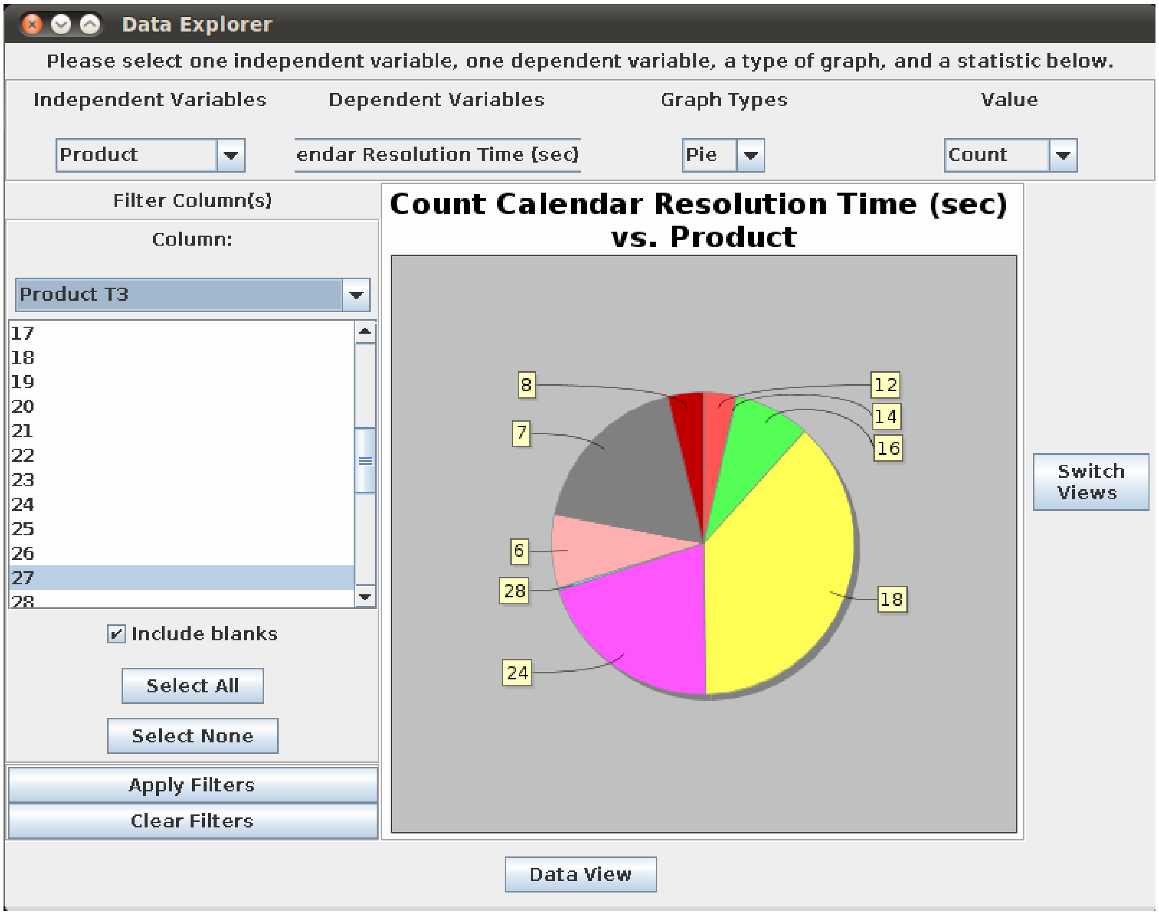
Independent Variables	Dependent Variables	Graph Types	Value
Product T3	endar Resolution Time (sec)	Pie	Count

Filter Column(s)

Column: Product T2

☑ Include blanks

Fig. 33. Count of Incidents by Product T3 - Pie Chart



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables	Dependent Variables	Graph Types	Value
Product	endar Resolution Time (sec)	Pie	Count

Filter Column(s)

Column: Product T3

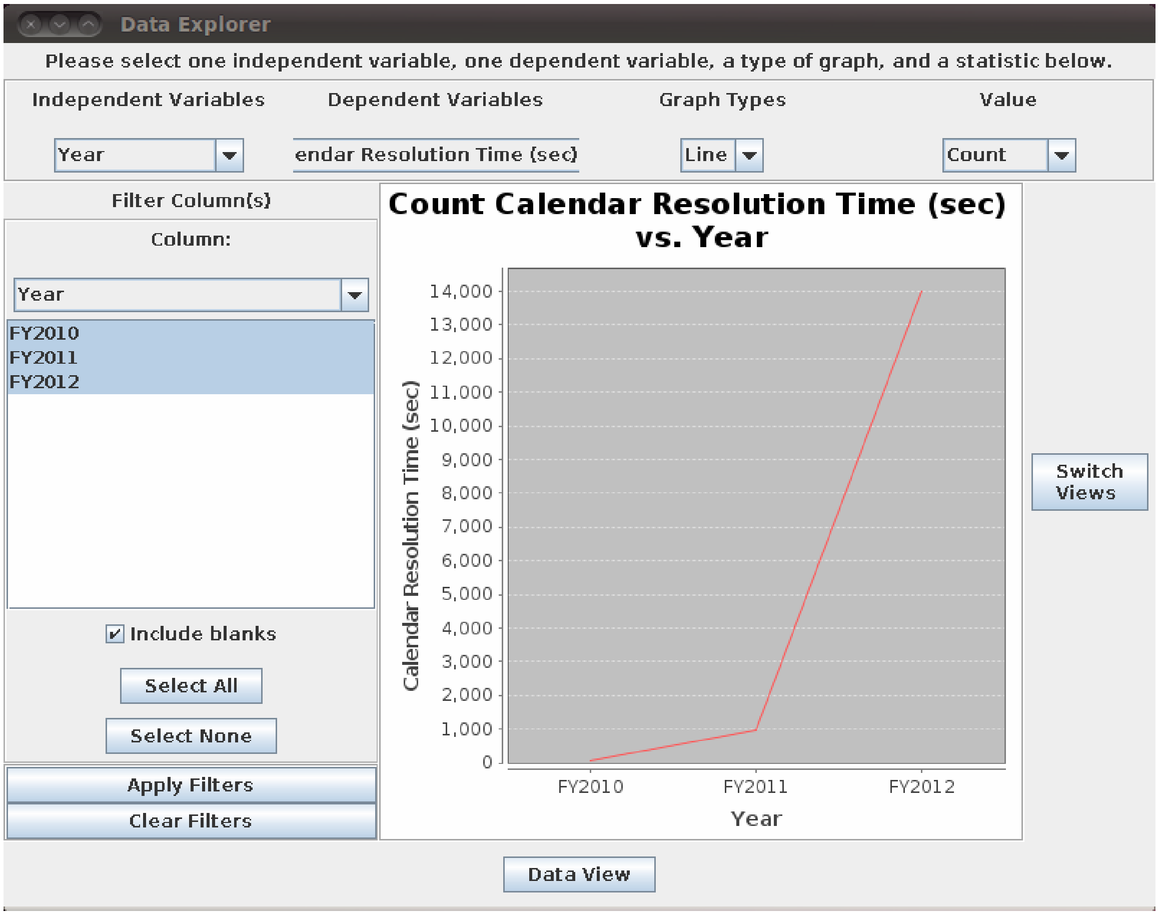
17
18
19
20
21
22
23
24
25
26
27
28

☑ Include blanks

Fig. 34. Count of Incidents by Product - Pie Chart

Incident Counts by Date

Another important grouping of the number of incidents is by date. Incidents are logged at a very high rate. Monitoring the number of incidents created over time is important for managing workloads for support teams. Figures 35 through 39 show the progression of line charts with more specific levels of filters, moving from the year, to the quarter, to the month, to the week, down to the day. The graphs are filtered on the previous category which was the largest number of incidents. An additional area for investigation identified by these charts is the volume of incidents based on the day of the week, the week of the month, the month of the quarter, and the quarter of the year. It appears that there is a significant difference between those values for each category. The sample data does not have columns to allow this grouping, but columns could be derived from existing columns.


Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Year

Dependent Variables: endar Resolution Time (sec)

Graph Types: Line

Value: Count

Filter Column(s)

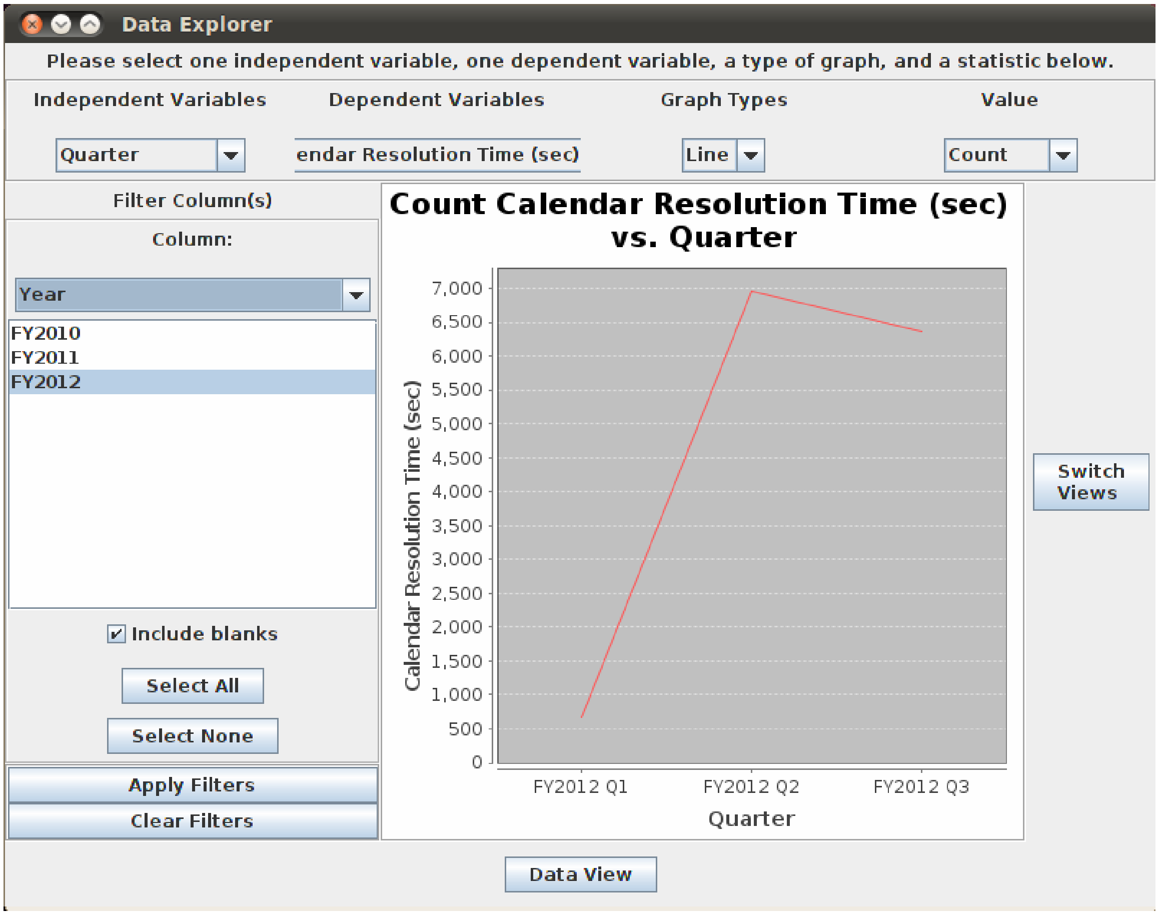
Column: Year

FY2010
FY2011
FY2012

☑ Include blanks

Fig. 35. Count of Incidents by Year – Line Chart

Figure 35 shows the majority of the incidents in the sample data occurred in FY2012, cuing the user to change the year filter to FY2012 and the independent variable to Quarter, as shown in Figure 36.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Quarter

Dependent Variables: endar Resolution Time (sec)

Graph Types: Line

Value: Count

Filter Column(s)

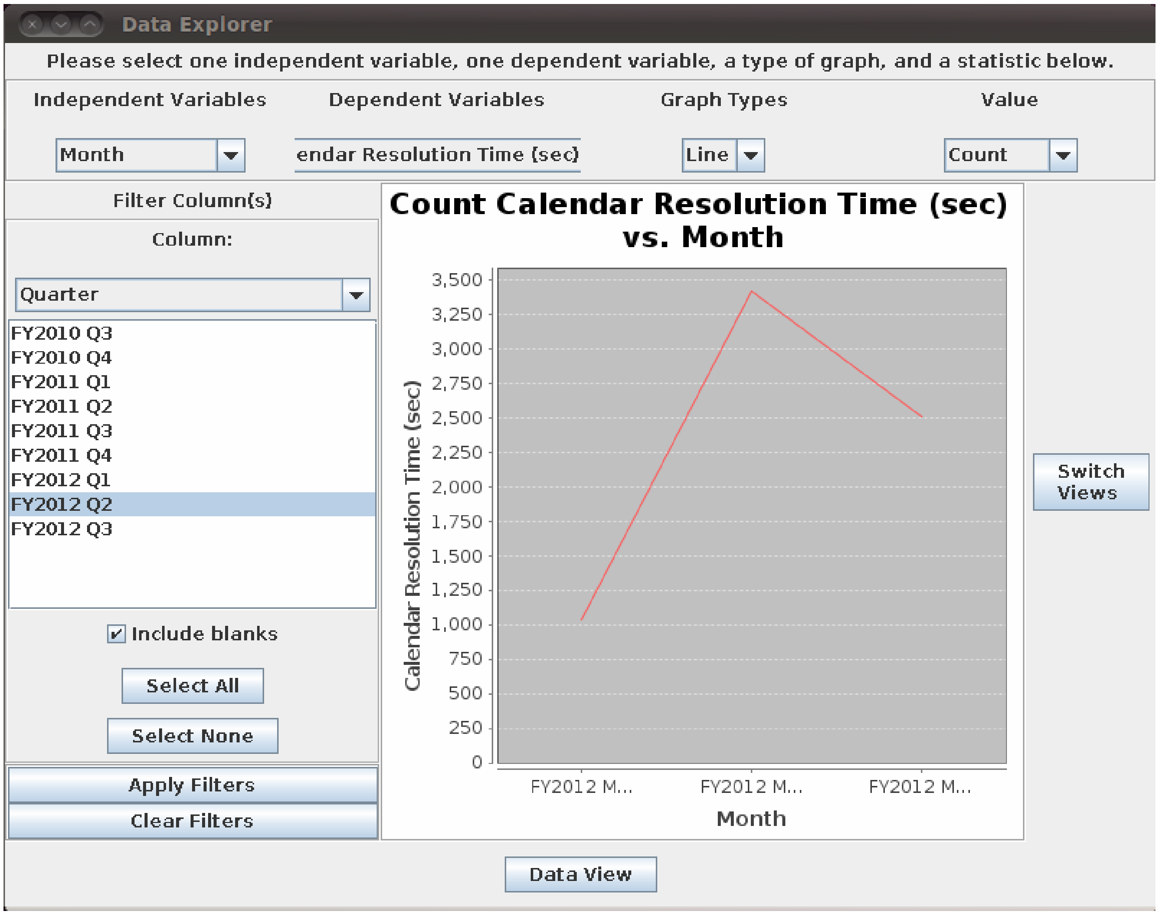
Column: Year

FY2010
FY2011
FY2012

☑ Include blanks

Fig. 36. Count of Incidents by Quarter - Line Chart

The second quarter of FY2012 has the largest number of incidents, seen in Figure 36. The filter is adjusted to focus on FY2012 Q2 with the independent variable set to Month as shown in Figure 37.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Month

Dependent Variables: endar Resolution Time (sec)

Graph Types: Line

Value: Count

Filter Column(s)

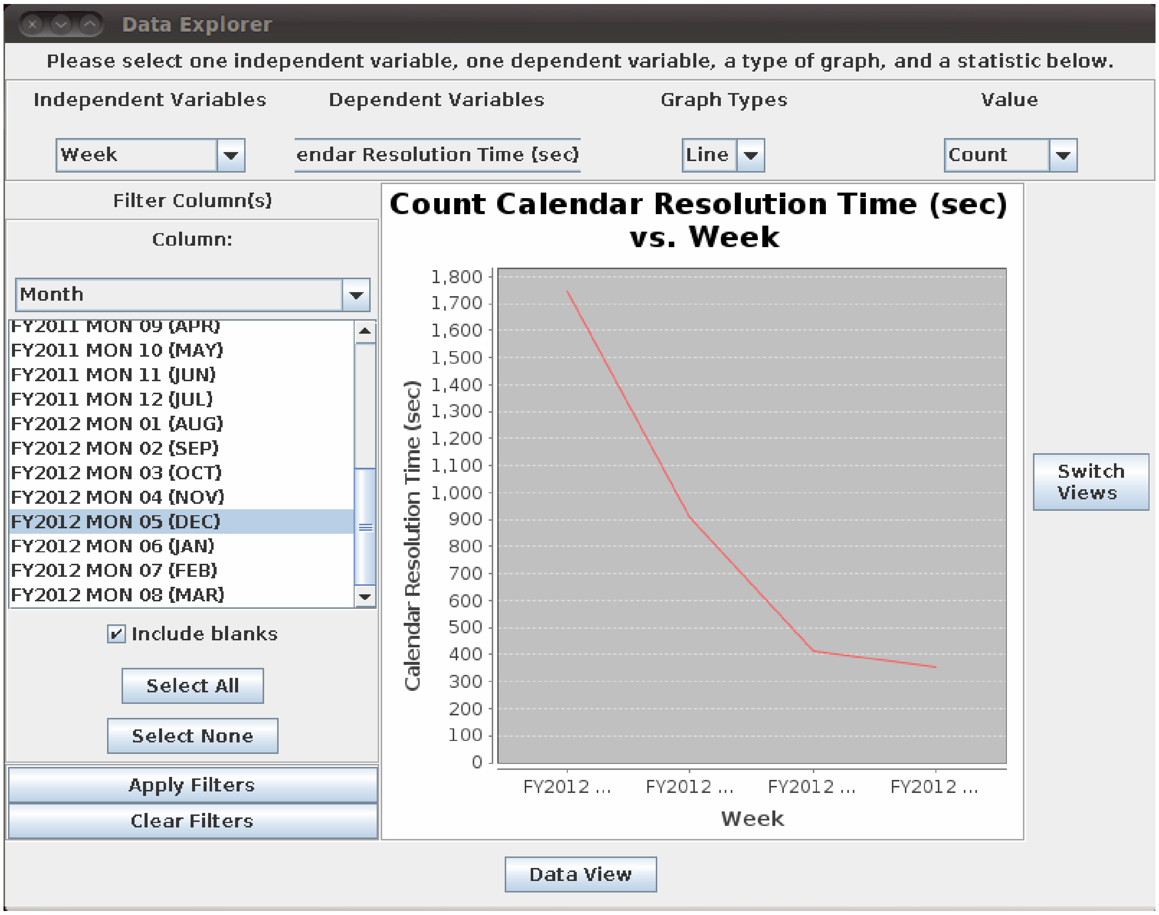
Column: Quarter

FY2010 Q3
FY2010 Q4
FY2011 Q1
FY2011 Q2
FY2011 Q3
FY2011 Q4
FY2012 Q1
FY2012 Q2
FY2012 Q3

☑ Include blanks

Fig. 37. Count of Incidents by Month - Line Chart

The second month of the quarter has the highest number of incidents. The filter is adjusted to focus on FY2012 MON 05 (DEC) and the independent variable to Week, as shown in Figure 38.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Week

Dependent Variables: endar Resolution Time (sec)

Graph Types: Line

Value: Count

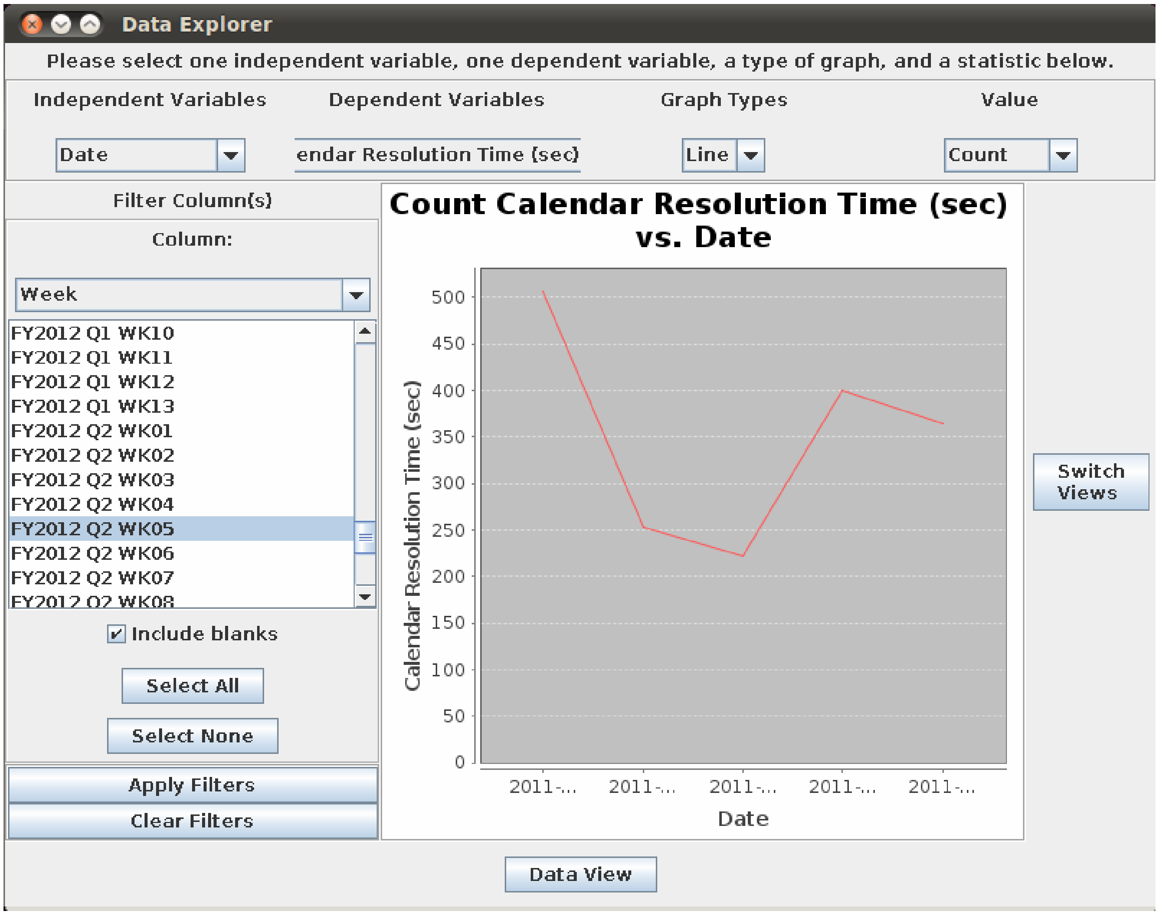
Filter Column(s)

Column: Month

☑ Include blanks

Fig. 38. Count of Incidents by Week - Line Chart

The number of incidents reported decreases significantly across December in 2012. This is a direct result of the year-end shutdown implemented by the company. The number of incidents reported in the first week of the month is still very high, over 10% of the total sample of incidents. The filter is changed to look at FY2012 Q2 WK05 and the independent variable to Date, as shown in Figure 39.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Date

Dependent Variables: endar Resolution Time (sec)

Graph Types: Line

Value: Count

Filter Column(s)

Column: Week

FY2012 Q1 WK10
FY2012 Q1 WK11
FY2012 Q1 WK12
FY2012 Q1 WK13
FY2012 Q2 WK01
FY2012 Q2 WK02
FY2012 Q2 WK03
FY2012 Q2 WK04
FY2012 Q2 WK05
FY2012 Q2 WK06
FY2012 Q2 WK07
FY2012 Q2 WK08

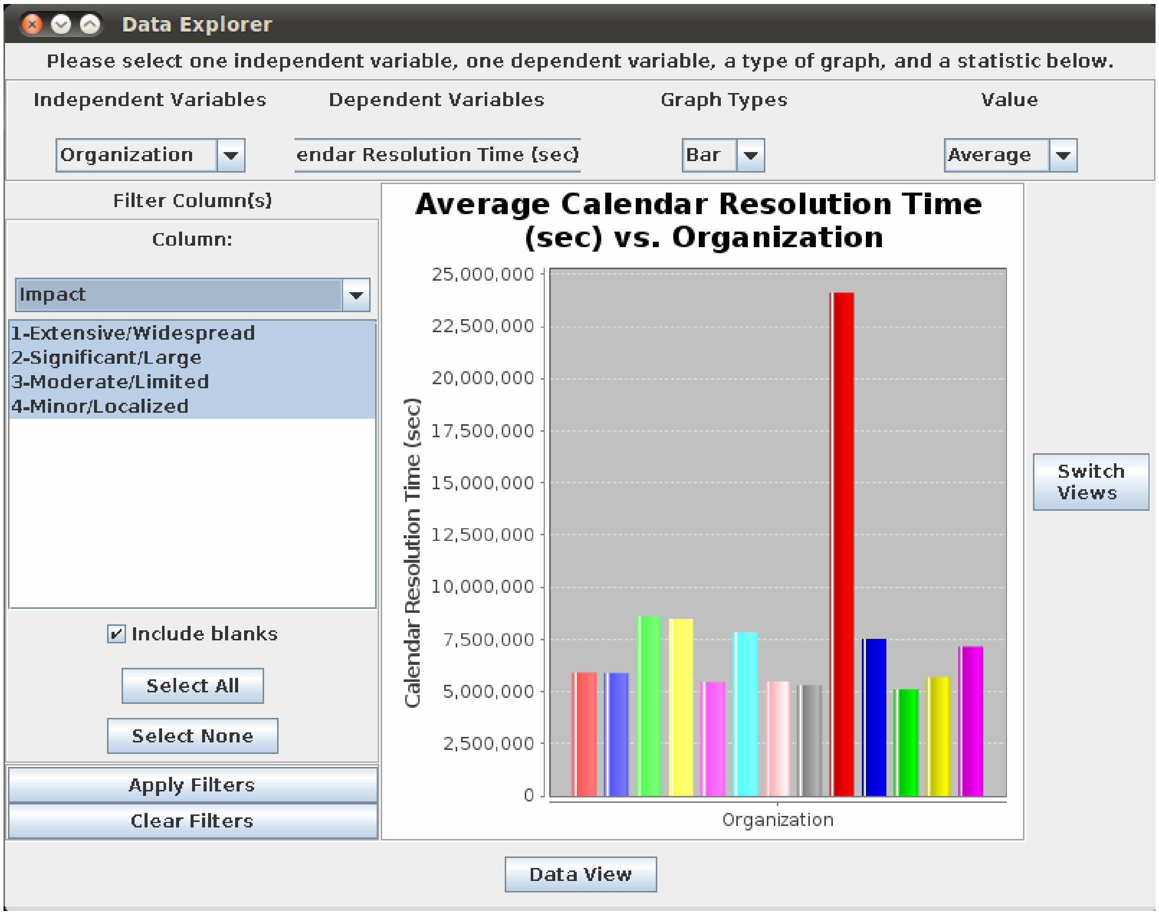
☑ Include blanks

Fig. 39. Count of Incidents by Day - Line Chart

Further investigation is needed to determine why volumes of incidents vary from day-to-day within the week in Figure 39. The analysis could be performed to look at the day within the week for all weeks throughout the year.

Calendar Resolution Time by Organization

Not only should the number of incidents for each product and the number of incidents created over time be monitored, but the performance of teams who are working to resolve the issues is critical to the success of the company. Figures 40 and 41 show bar charts with the average resolution time and the count of incidents for Organizations. Interestingly, Organization 4 (in red) has the highest resolution time shown in Figure 40, while Organization 2 (in cyan) has the highest number of incidents shown in Figure 41.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Organization

Dependent Variables: endar Resolution Time (sec)

Graph Types: Bar

Value: Average

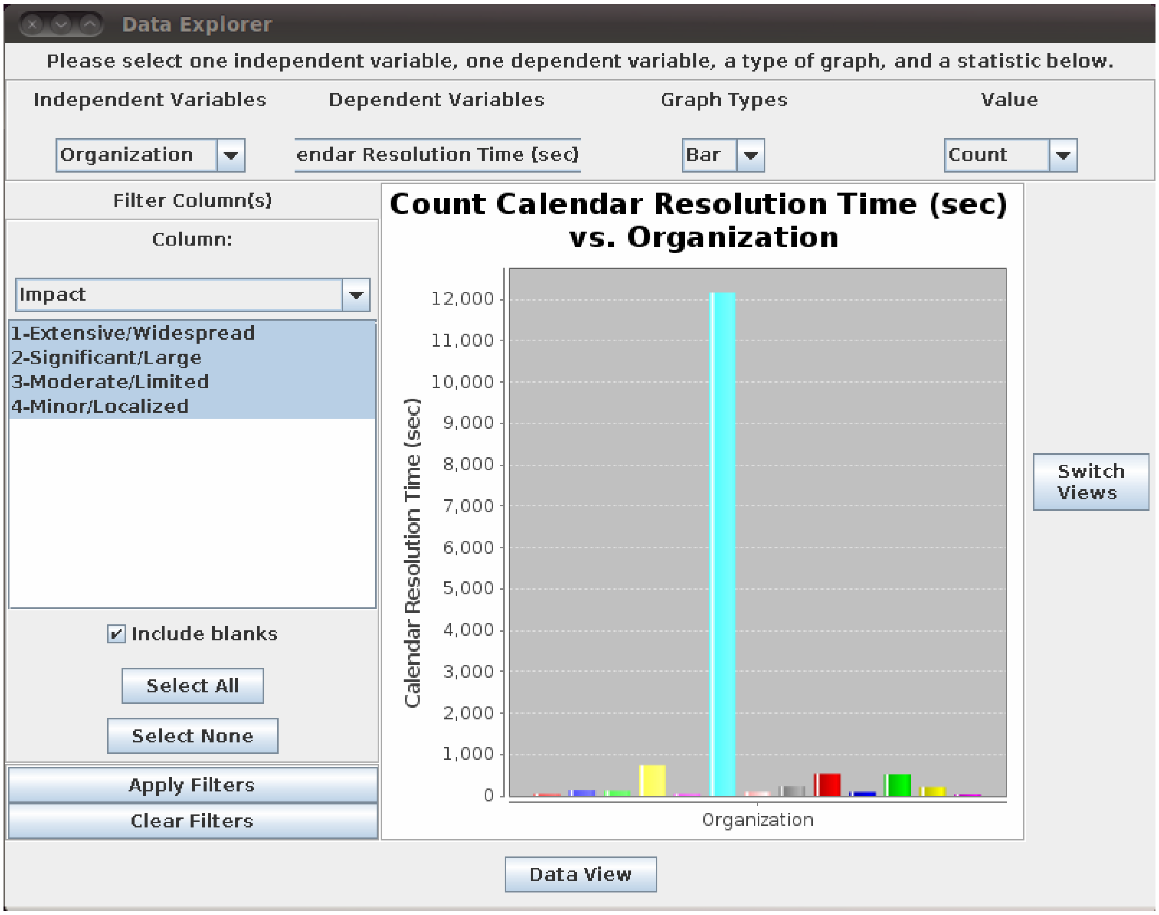
Filter Column(s)

Column: Impact

1-Extensive/Widespread
2-Significant/Large
3-Moderate/Limited
4-Minor/Localized

☑ Include blanks

Fig. 40. Average Resolution Time by Organization - Bar Chart



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Organization

Dependent Variables: endar Resolution Time (sec)

Graph Types: Bar

Value: Count

Filter Column(s)

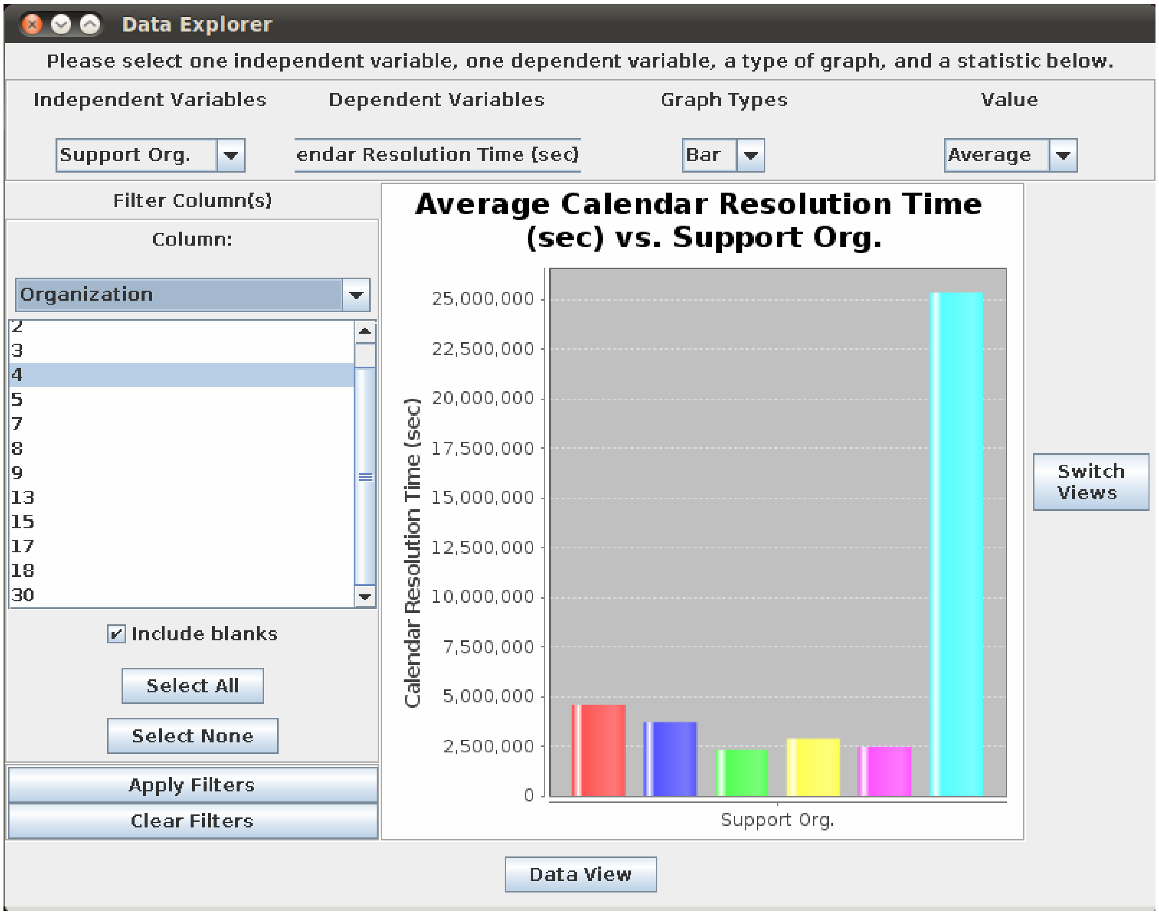
Column: Impact

1-Extensive/Widespread
2-Significant/Large
3-Moderate/Limited
4-Minor/Localized

☑ Include blanks

Fig. 41. Count of Incidents by Organization - Bar Chart

Figures 42 and 43 contain bar charts with the average resolution time and the count of incidents for Support Orgs belonging to Organization 4. Support Org. 9 is receiving over 99% of the total incidents, which appears to be impacting the resolution times as well.



Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables	Dependent Variables	Graph Types	Value
Support Org.	endar Resolution Time (sec)	Bar	Average

Filter Column(s)

Column:

Organization

2
3
4
5
7
8
9
13
15
17
18
30

☑ Include blanks

Fig. 42. Average Resolution Time by Support Org. in Organization 4 - Bar Chart

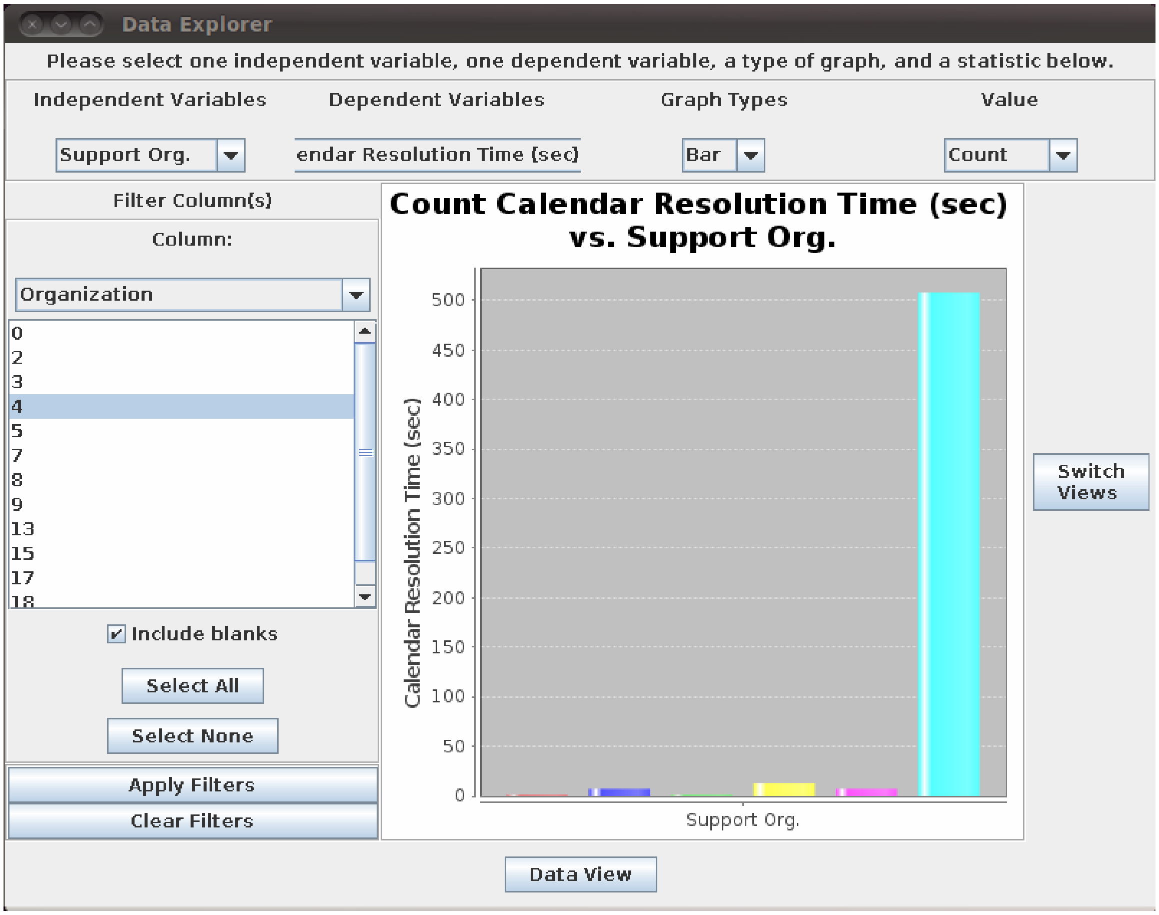
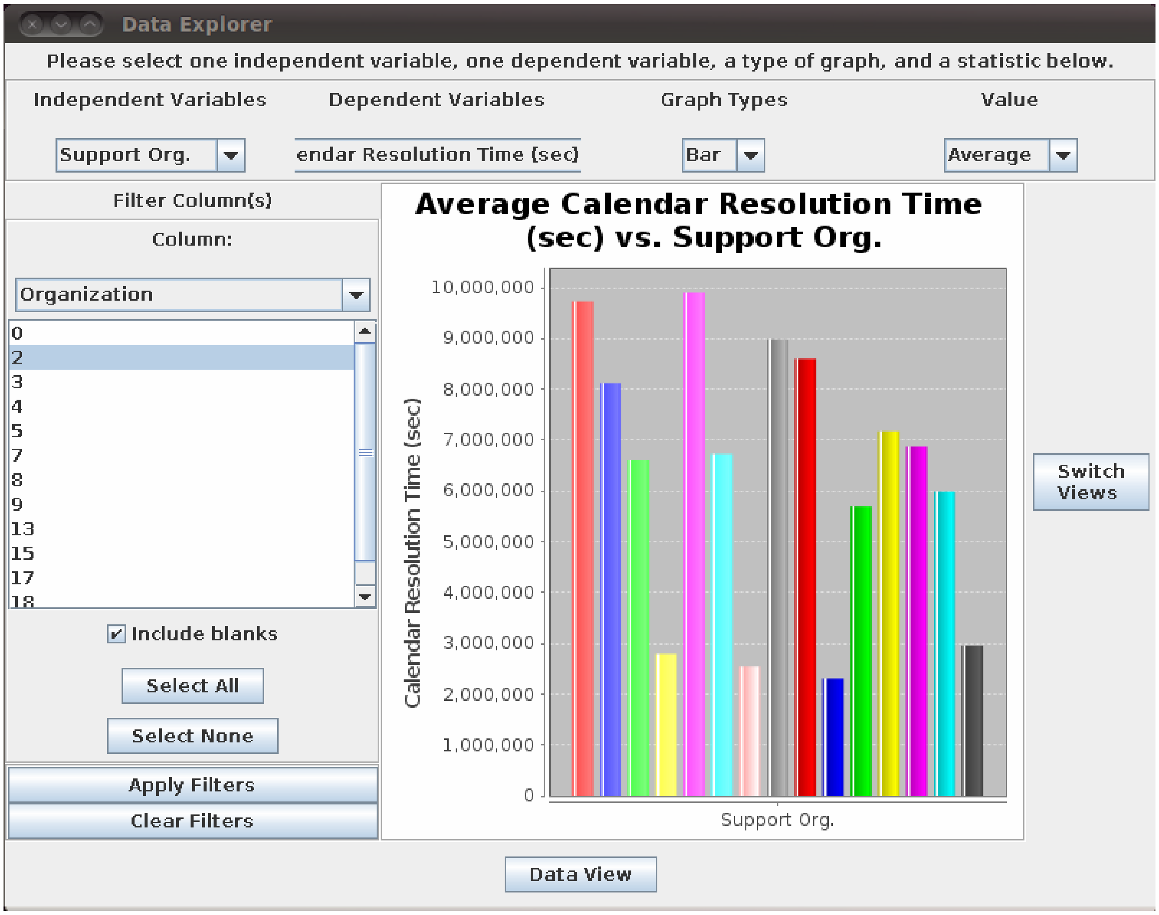


Fig. 43. Count of Incidents by Support Org. in Organization 4 - Bar Chart

Figures 44 and 45 contain bar charts with the average resolution time and the count of incidents for Support Orgs belonging to Organization 2. Figure 44 shows the average resolution time is widely distributed between the Support Org's but no particular Support Org. stands out. This is even more surprising when Figure 45 shows Support Org. 15 (in blue) has high volume of incidents.



Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables	Dependent Variables	Graph Types	Value
Support Org.	endar Resolution Time (sec)	Bar	Average

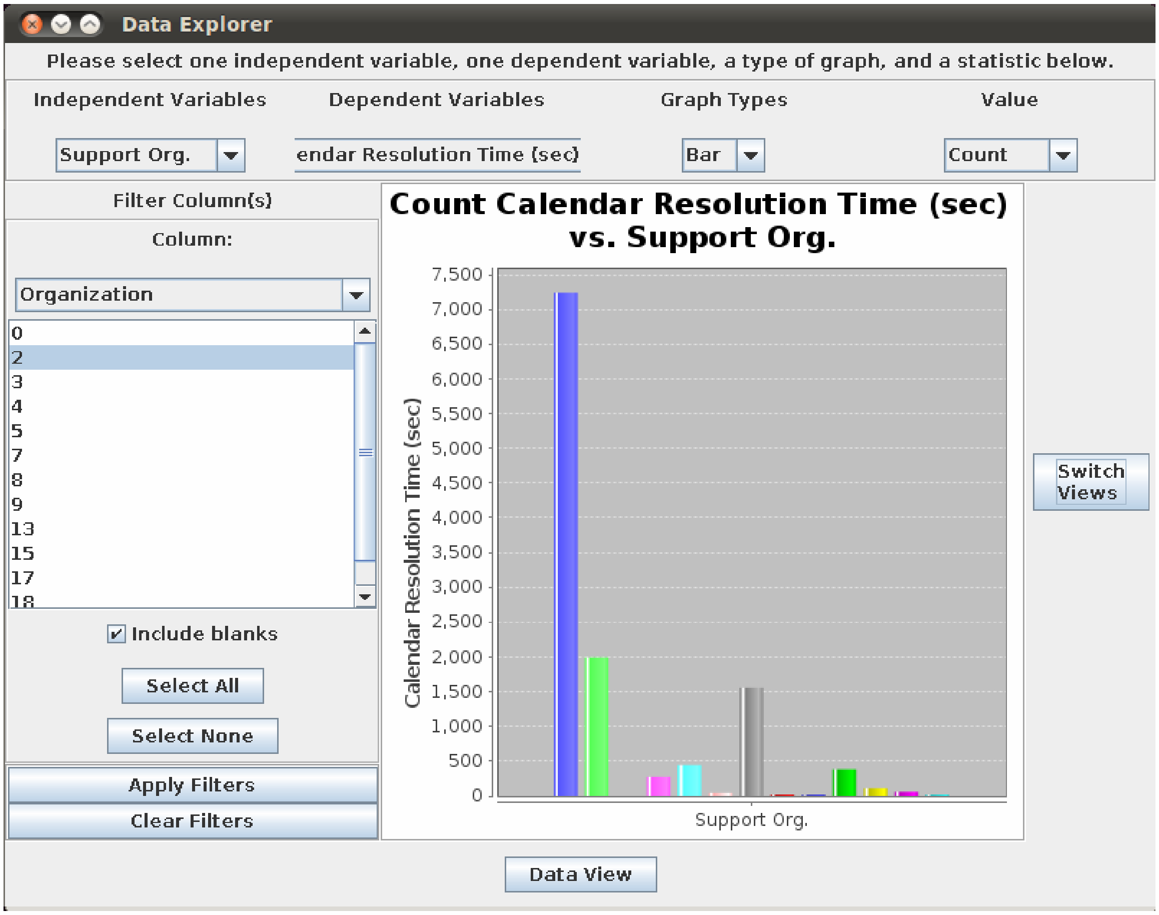
Filter Column(s)

Column: Organization

0
2
3
4
5
7
8
9
13
15
17
18

☑ Include blanks

Fig. 44. Average Resolution Time by Support Org. in Organization 2 - Bar Chart



Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables: Support Org.

Dependent Variables: endar Resolution Time (sec)

Graph Types: Bar

Value: Count

Filter Column(s)

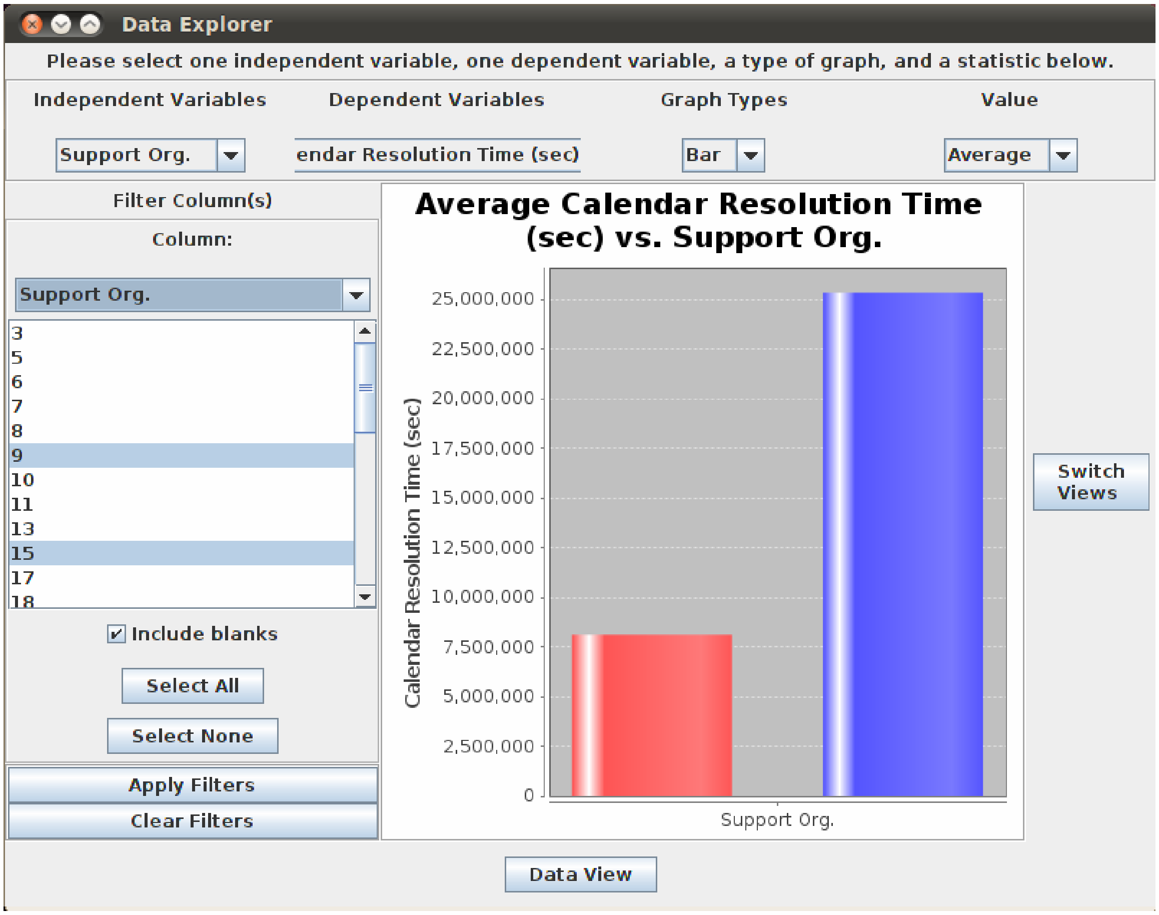
Column: Organization

0
2
3
4
5
7
8
9
13
15
17
18

☑ Include blanks

Fig. 45. Count of Incidents by Support Org. in Organization 2 - Bar Chart

Figures 46 and 47 contain bar charts with the average resolution time and the count of incidents for Support Org. 9 (in blue) and 15 (in red). Figure 46 shows the average resolution time for Support Org. 9 is nearly three times that of Support Org. 15. Figure 47 shows the count of incidents for Support Org. 9 is roughly 500 incidents, while Support Org. 15 exceeds 7000 incidents, close to 50% of the sample data.



Data Explorer

Please select one independent variable, one dependent variable, a type of graph, and a statistic below.

Independent Variables	Dependent Variables	Graph Types	Value
Support Org.	endar Resolution Time (sec)	Bar	Average

Filter Column(s)

Column: Support Org.

3
5
6
7
8
9
10
11
13
15
17
18

☑ Include blanks

Fig. 46. Average Resolution Time for Support Orgs 9 & 15 - Bar Chart

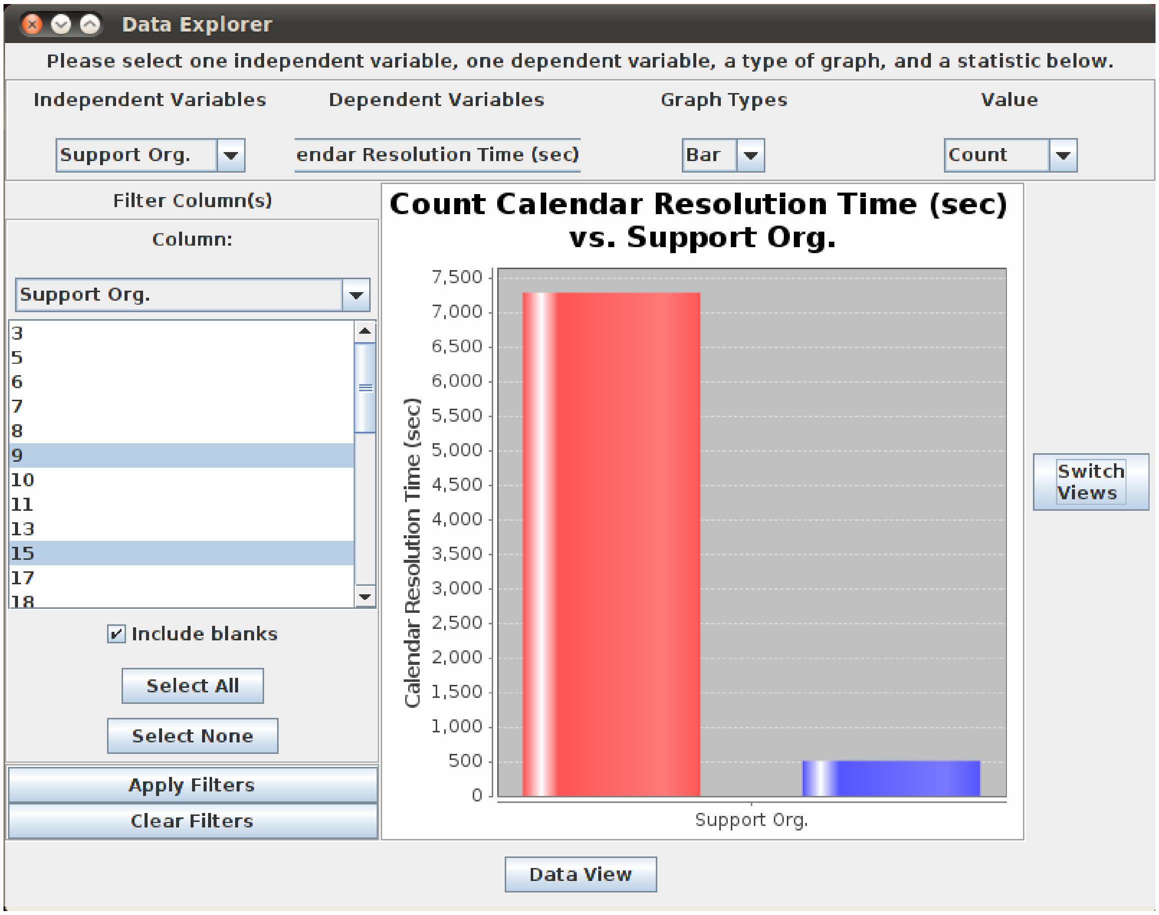


Fig. 47. Count of Incidents for Support Orgs 9 & 15 - Bar Chart

One note is the average calendar resolution times in Figure 46 are rather large. The average calendar resolution time for Support Org. 9 is close to ¾ of a year. Incidents that are not closed quickly are often lost in the queue of incoming work. Both Support Orgs should review their open incidents to check for incidents which can be closed.

Best practices from Support Org. 15 should be shared with Support Org. 9 to reduce resolution times. Support Org. 9 may also have too much work given the type of incidents being resolved, requiring further delegation and distribution of the work.

CHAPTER 5

Future Work

The methodology proposed focuses on improving the data exploration process for any user on a desktop computer. The scope was limited to improve the focus of the research and to fit within the available timeframe. There are additional areas which could be explored to further improve the data exploration process discussed below.

Web Availability

A software tool which implements this methodology is ideally suited to be made available through a web page. Additional development of the Java implementation could utilize a Java Applet to be embedded directly into a web page. The page could require authentication and also allow users to store data and results on a secured server for retrieval and information sharing. The software-as-a-service (SaaS) model could be used to work with a wide variety of organizations.

Database Support

Although a stand-alone application reduces the amount of systems integration, many of the steps in the methodology can be implemented much more efficiently using an application in conjunction with a database management system. Client side database management systems such as Oracle's MySQL or Microsoft Access could be used for a single user who does not share data. The main obstacle could be the performance decrease of the database software while responding to requests from an additional user application.

Expanded Data Mining Incorporation

The purpose of this methodology is to assist with the data exploration phase of data mining, but data mining algorithms including regression analysis and cluster analysis could be automatically applied to data sets using the information gained using the previous steps of the methodology. The methodology could be expanded to include those analyses, as well as others, to further automate the process and alleviate the time spent by an individual user.

One caution in adding these additional analyses is that the performance of any implementation of the methodology will decrease. Data mining analysis algorithms are powerful but require computing power to work effectively. Automatically performing all possible analyses may also be undesirable from this standpoint. The time to results will increase exponentially.

CHAPTER 6

Conclusion

The author presented a methodology to allow data exploration to be performed quickly and easily. The methodology incorporates visualizations into data exploration and automates their creation to further simplify the task for the user.

The methodology and prototype implementation show that the automation and improved standardization of the data exploration phase of data mining in support of KDD or analytics can reduce the effort of individual analysts and the time required to produce useful information. The methodology also allows the analyst to gain perspective on the data from multiple levels, ranging from the entire data set to a very detailed subset. A standardized methodology will have benefits across a variety of fields, especially those without the knowledge of high performance computing systems.

Big Data is an emerging issue and will continue to become relevant in an increasing number of fields. Additional methodologies to further standardize the KDD and analytics processes will be needed. Methodologies and tools do not need to be complex. Such tools should focus on the analyst or user who will be following the methodology rather than on the volume or complexity of the data.

The two cases presented immediately show the value of the author's methodology. The reduction in time required to produce useful information from data sets will allow more data to be reviewed. The untapped potential of volumes of stored data can gradually be unlocked. More importantly, the methodology provides the information about data needed to determine the most effective method of data mining for use in KDD or analytics. Organizations following the methodology will benefit regardless of their experience with analytics, and less experienced organizations will be able to use the methodology to assess where analytics will be most effective.

LIST OF REFERENCES

[1] Al Ghoson, Abdullah M. "Decision tree induction & clustering techniques in SAS Enterprise Miner, SPSS Clementine, and IBM Intelligent Miner – a comparative analysis". *International Journal of Management and Information Systems* 14.3 (2010): 57-70. ABI/INFORM Complete. Web. 10 February 2013.

[2] Babu, G. Jogesh,. and Feigelson, Eric D. "Big data in astronomy". *Significance* 9.4 (2012): 22-25. Wiley Online Library. 5 February 2013.

[3] Ball, Phillip. "Data visualization: Picture this". *Nature* 418.6893 (2002): 11-13. Nature. Web. 10 February 2013.

[4] Batty, Michael. "Big data; big issues". *Geographical* 85.1 (2013): 75. Gale Cengage Academic OneFile. Web. 5 February 2013.

[5] Bose, Ranjit. "Advanced analytics: Opportunities and challenges". *Industrial Management and Data Systems* 109.2 (2009): 156-172. Emerald. Web. 18 October 2012.

[6] Bremer, Stuart., Singer, J. David,. and Stuckey, John. "Capability distribution uncertainty, and major power war, 1820-1965". *Peace, War, and Numbers* (1972): 19-48. Print.

[7] Brodley, Carla E., Lane, Terran., and Stough, Timothy M. "Knowledge discovery and data mining: Computers taught to discern patterns, detect anomalies and apply decision algorithms can help secure computer systems and find volcanoes on Venus". *American Scientist* 87.1 (1999): 54-61. JSTOR. Web. 5 February 2013.

[8] Bullard, James., Dudoid, Sandrine., Durink, Steffen., and Spellman, Paul T. "GenomeGraphs: Integrated genomic data visualization with R". *BMC Bioinformatics* 10.2 (2009) Gale Cengage Academic OneFile. Web. 10 February 2013.

[9] Chabot, Christian. "Demystifying visual analytics". *IEEE Computer Graphics and Applications* 29.2 (2009): 84-87. IEEE Xplore Journals. Web. 18 October 2012.

[10] Chen, Chaomei., Johnson, Christopher R., Ross, Robert B., Shen, Han-Wei., and Wong, Pak Chung. "The top 10 challenges in extreme-scale visual analytics". *IEEE Computer Graphics and Applications* 32.4 (2012): 63-67. IEEE Xlpore Journals. Web. 25 October 2012.

[11] Cook, Kristin A., and Thomas, James J. "A visual analytics agenda". *IEEE Computer Graphics and Applications* 26.1 (2006): 10-13. IEEE Xplore Journals. Web. 18 October 2012.

[12] Cook, Kristin A., and Thomas, James J. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos: IEEE Computer Society Press, 2005. Print.

LIST OF REFERENCES (continued)

[13] *Correlates of War Project National Materials Data Documentation Version 4.0*. 2010. Correlates of War. Web. 30 January 2013.

[14] Courtney, Martin. “Puzzling out big data”. *Engineering & Technology* 7.12 (2012): 56-60. IEEE Xplore Journals. Web. 5 February 2013.

[15] Fayyad, Usama., Piatetsky-Shapiro, Gregory., and Padhraic, Smyth. “Knowledge discovery and data mining: Towards a unifying framework”. *Second International Conference of Knowledge Management and Data Mining* 2.6 (1996): 82-88. Association for the Advancement of Artificial Intelligence. Web. 18 October 2012.

[16] Fayyad, Usama., Piatetsky-Shapiro, Gregory., and Padhraic, Smyth. “The KDD process for extracting useful knowledge from volumes of data”. *Communications of the ACM* 39.11 (1996): 27-34. ACM Digital Library. Web. 5 February 2013.

[17] Finch, Glenn., Haydock, Michael., Kiron, David., Kruschwitz, Nina., and Shockley, Rebecca. “Analytics: The widening divide”. *MIT Sloan Management Review* 53.2 (2012): 1-21. ABI/INFORM Complete. Web. 18 October 2012.

[18] Goldston, David. “Data wrangling”. *Nature* 455.7209 (2008): 15. Gale Cengage Academic OneFile. Web. 5 February 2013.

[19] Hen, Lai Ee., and Lee, Sai Peck. “Performance of data mining tools cumulating with a proposed data mining middleware”. *Journal of Computer Science* 4.10 (2008): 826-833. Gale Cenage Academic OneFile. Web. 10 February 2013.

[20] Hoffman, Leah. “Looking back at big data”. *Communications of the ACM* 56.4 (2013): 21-23. ACM Digital Library. Web. 15 April 2013.

[21] *Human Genome Project Information*. U.S. Department of Energy Office of Science, 31 July 2012. Web. 5 February 2013.

[22] Jackson, Russell A. “Big data”. *Internal Auditor* 70.1 (2013): 34-38. Gale Cengage Academic OneFile. Web. 25 March 2013.

[23] Jacobs, Adam. “The pathologies of big data”. *Communications of the ACM* 52.8 (2009): 36-44. ACM Digital Library. Web. 5 February 2013.

[24] Kiron, David., and Shockley, Rebecca. “Creating business value with analytics”. *MIT Sloan Management Review* 53.1 (2011): 57-63. ABI/INFORM Complete. Web. 18 October 2012.

[25] Kobielus, James. “The Forrester Wave: Predictive analytics and data mining solutions, Q1 2010”. *Forrester: For Business Process & Applications Professionals* (2010): 1-20. IBM. Web. 18 October 2012.

[26] Leventhal, Barry. “An introduction to data mining and other techniques for advanced analytics”. *Journal of Direct, Data and Digital Marketing Practice* 12.2 (2010): 137-153. ABI/INFORM Complete. Web. 18 October 2012.

[27] Leventhal, Barry. “Leveraging the census for customer analysis”. *Journal of Targeting, Measurement and Analysis for Marketing* 12.1 (2003): 11-19. ABI/INFORM Complete. Web. 18 October 2012.

[28] Lemieux, Victoria L. “Visual analytics: A new way to manage data deluge in e-discovery”. *Information management Journal* 45.2 (2011): 38-40. ABI/INFORM Complete. Web. 18 October 2012.

[29] Purohit, Neha., Purohit, Sapna., and Purohit, Ritesh Kumar. “Data mining applications and knowledge discovery”. *International Journal of Advanced Computer Research* 2.6 (2012): 458-462. The Accents. Web. 5 February 2013.

[30] *Rapid-I*. Rapid-I. Web. 12 February 2013.

[31] Rose, Sherri. “Big data and the future”. *Significance* 9.4 (2012): 47-48. Wiley Online Library. 5 February 2013.

[32] Singer, J. David. “Reconstructing the correlates of war dataset on material capabilities of states, 1816-1985” *International Interactions* 14 (1987): 115-132. Taylor & Francis Online. Web. 30 January 2013.

[33] “The power of big data”. *The UMTRI Research Review* 43.2 (2012): 2-3. ProQuest. Web. 5 February 2013.

[34] Van der Meulen, Rob., and Petty, Christy. “Gartner says more than 1 billion PCs in use worldwide and headed to 2 billion units by 2014”. *Gartner.com*. Gartner, Inc. 23 June 2008. Web. 2 February 2013.

[35] “What is big data?”. *Autonomy.com*. Hewlett Packard Autonomy. 8 August 2012. Web. 2 Feburary 2013.

[36] “What is the cost of SAS Enterprise Miner & SPSS Clementine?”. *Discussions: Advanced Business Analytics, Data Mining and Predictive Modeling*. LinkedIn, 13 June 2011. Web. 10 February 2013.