Visualization and Data Analysis
Spend-O-Tron

Artan Toplanaj*
University of Vienna
1068861

Dominic Palffy†
University of Vienna
01302603

Jiri Mauritz‡
University of Vienna
11772172

1 Motivation

Our project is called Spend-O-Tron and its focus is the USA government spending data. We wanted to have a closer look to the spending habits of the government. It is always intriguing to know how the money of a government is distributed, in this way we can confirm the priorities of that government and thus also understand its politics and its decision in different situations.

The data used can be found on `https://www.usaspending.gov`. There can be found 4 different spending types include Contracts, Grants, Loans and Other Financial Assistance. Since most of the money is distributed for Contracts, we decided that this data set is the one we would like to take a closer look at. Since the data set for 2017 wasnt finished yet, we decided to use the data from 2016.

This data is so huge, that without the right tool, a user, that wants to know something more about the spending habits of its government, cannot come up with any conclusions! This was one problem with the data. It needed also quite some time to just load the data. This is because the data set holds lots of information about every transaction, including things like Department, Agency, Product, Base Amount, Requested Amount, Address, Date and lots of other information about the vendors, like Vendor Name, Number of Employees, Address etc. The data set has more than 4 million rows!

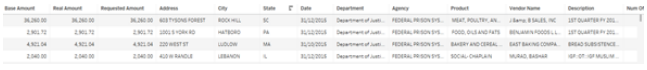

Figure 1: First view of the data

When the data is this big, there are of course lots of unneeded information, which had to be filtered out. Filtering the data was one of the biggest challenges. There was lots of information that had to be filtered out, but the kernel of the data set should be untouchable. This was very time consuming!

The idea of our team was to create a tool that any user could use, regardless of his background. The user could be an entrepreneur who wants his business founded, a farmer who wants to find any evidence of mishandled funds from the government, the head of an established company doing contract work for the government or even insiders from the government who want to see the money distribution and would like also more information about the vendors.

Our tool should provide a good overview of the most important data, like the top receiving departments, the top receiving agencies in that department, top vendors, the regions that were receiving most of the money. Other than that, our tool should make it possible to the user to click around over our views and to get the information he was looking for. Thus, the user could accomplish different task with our tool, like he could find out the amount of money given to companies by the government and the date of the transfer. The user should also find out about the general distribution of the money and which departments where receiving most of the money. Departments like military, education, healthcare etc. should be able broken down into their agencies for more detailed information in that particular sector. A spending map should help the user to find out which regions receive the most money from the budget of the government. Another task that the user should be able to do, would be showing government spending over specific periods of time, to point out the reactions of government in case of natural disasters. The user should be also able to take a closer look at the suspicious transactions to find correlations which may point to illegal behaviour.

*e-mail: artantoplanaj@gmail.com
†e-mail: dominicpalffy@gmail.com
‡e-mail: jirmauritz@gmail.com

2 Related Work

One similar approach was made from UsaSpending.gov and it can be found on `https://beta.usaspending.gov/#/`. Its a very nice tree map combined with a simple heatmap, that has some colour encoding and a tooltip.

The tree map visualizes the biggest spenders. From the tree map we can see that 3 of the 19 total budget functions are accounted for about of total spending. These so-called functions are Social Security, National Defense and Medicare. This tree map is interactive, so the user can click on one of the functions and underneath is going to be shown the Departments in these functions. On the tooltip is included the name of the Department, the amount of money spent and its percentage compared to other departments of the same function and a short description.

These two techniques (tree map and heatmap) can be found also in our visualization tool, but we didnt incorporate these two techniques from this approach. The information we got from the class and knowing that there is a hierarchy in the data, led us to using the heat map. If in the data there is a zip code, that you can project in the map, then most of the time a heat map is going to be a natural choice for that.

To visualize our tool, we have used Tableau 10.4 with a student licence. Tableau was introduced to us in the Vis-class and it offered us everything we needed for our visualization, thats why we decided for that. Its easy very easy to use and we think we are going to be using Tableau in the future as well!

3 Approach

Our final working set of attributes contains transaction amount (base, real and requested), address, date, department, agency, product, and vendor name. We would like to provide a user with the ability to research any group of transactions filtered based on category, location and time. Therefore, we designed a dashboard, where the top part serves as a specification and localization controller and the bottom part serves as the research part.

3.1 Category

The category filtering was driven by the hierarchy in the following attributes:

1. Department - major federal organizations.

2. Agency - governmental agency or bureau.

3. Product - what type of goods appear as the subject of the transaction.

4. Vendor - the company which received money.

As our first idea, we considered a clickable treemap which shows the next level of hierarchy as a user clicks through the department, agency or product. According to the feedback, we realized that it might not be the best idea since we have just a few levels.

However, we still want a user to understand the hierarchy and provide her with the possibility to explore it. The final treemap contains only the first two levels (departments and agencies) where the separated rectangles represent agencies and departments are encoded by color. Treemap structure does not change by any interaction so that the user see which department is active all the time. Interaction with the treemap highlights the correct agencies and filters records in other charts. We decided not to display small departments since the treemap is very bad at displaying proportionally tiny values. More importantly, we are generally focusing on the major subjects in the overall dashboard because our use cases showed that a user rarely needs to see the minor ones.

The remaining two levels of hierarchy deserve its own graphs. The product is visible in the Drill Down Bar, which displays the top five products for the selected agencies. The color helps users to link the selection to the relating department.The Drill Down Bar starts to be unreadable with a large number of products, therefore we focus only on the top five. By the aforementioned majority rule, we expect user not to be concerned with minor products.

Regarding the vendors, we present a single bar chart displaying the sorted total amount of transactions per vendor. Extra information hidden in this chart is also how much money goes to the specific company from each department, which is encoded by colored bars, similarly as in the case of product bars (however, there each bar is single-colored).

3.2 Location

The dataset contains information about the recipients address including state, city and zip code. For simplicity, we have decided to aggregate the data per state and visualize the amount of money donated to each state by a heatmap. The value is the base amount, which is used in all category-related charts as well, not to confuse the user with options.

The heatmap is interactive, in a sense that user can pan, zoom and select by dragging the mouse across the wanted states. We picked an orange scale as a color pallet, which is not aggressive and is nice to perceive. The scale is adjusted every time a filtering is performed so that the highest value is always brown and the smallest beige. When a user hover over a state, a toolbox appears with information about how many vendors received money in this state and how much money it is. The heatmap is clickable and executes filtering by a state when clicked. Note that more states can be chosen when Ctrl is held (similarly as in the treemap).

3.3 Time

In the left bottom corner, we present an area chart displaying granted amount of money over a time period. The resolution of the time axis is at the level of months by default. However, a user can zoom out to the resolution of quartiles or years, or zoom in to the resolution of weeks or days. In the same manner, as in product/vendor bar charts, we included information about the types of departments in the area chart by color. Thanks to this feature, a user can discover replacing grants of one department with another in time, which would be impossible to discover in a chart of total amounts over all departments.

Filtering is possible by selecting a time period (quarter, month, week or day depending on the level of zoom). User can also pick a larger span by mouse selection or pick multiple distinct time periods, which enables maximal adjustability for a user.

3.4 Granted amounts

After the user is satisfied with the filtering, it is time to explore the transactions in the research part of the graph. We introduced two graphs for that purpose: Granted amounts per departments and Real vs. Requested amount scatter plot.

Granted amounts per departments are displayed in a multi-bar chart. For each selected department, there are three bars in total representing base, real and requested amount. The overall transaction amounts are easily comparable in this settings. We would expect the sum of real amounts to be less than the sum of requested and more than sum of the base amounts. In some cases, we can see that it is not that case and discover the irregularities in the grants.

Requested amount scatter plot also displays real and requested amounts but we do not include the base amount. Instead of departments, the data objects are vendors. The question we are asking here is: What companies receive less/more than they require?. To emphasize the purpose, we present the chart with the distinct color highlight of the vendors, for which the total sum of real amounts is larger, equal or smaller than the requested. We believe that the scatter plot can help to discover some outliers and strange occurrences of vendors regarding the granted amounts.

4 Implementation

For the preprocessing phase, we applied Unix tools and python scripting. Our original dataset contained 11 GB of data, which is not possible to process by standard editors. We took advantage of the python library Dask, which is designed for parallel analytical computing. It provides tools for working with a large structured data. We applied it to remove unnecessary attributes and faulty entries from the dataset. The final dataset is 814 MB large.

To implement our visualization we have decided to use Tableau. The decision was driven by the fact, that it already has the needed technology integrated for our task. In M1, we mentioned that we would use JavaScript for additional tasks not handled by Tableau, however, Tableau was able to provide sufficient level of solution for all of our requirements. We employed other convenient Tableau services, such as Tableau Server, Tableau Public, and Tableau Bridge. Since our dataset is rather large, we let the Tableau Server to accommodate the data and work with lightweight dashboard without data.

Challenges

One of our issues was dataset replacement while preserving all the charts. There were numerous versions of the dataset. One reason is that we have been often editing the attributes, and secondly, we worked with a subset of the full data at first and needed to switch to the original at the end. The replacement was impossible without rebuilding the charts from zero until we found a solution: upload the dataset to the Tableau Server and change links only.

Also, we hit one of the limitations of Tableau. One can set the actions to apply highlighting or selection to a chart. The problem is that we needed the chart to change the x-axis to contain another data field as a reaction to a selection in another chart. Such a behavior is possible with URL routing, however, it does not work in the desktop version but only on the server. Finally, we decided to change the design anyway, therefore we left this problem unsolved.

After we plugged in the full dataset, a problem with performance occurred. The response time of the filtering was quite high - more than 10 seconds. Generally, the filtering of larger categories took a bit longer and the filtering by states in the heatmap was rather slow. Hence, we reviewed the dataset and removed as much unnecessary

information as possible. For example, we also kept information about the number of employees of vendors and some descriptions about the transactions. After we dropped these, we achieved 5-7 seconds per filtering, which is acceptable for an ordinary user and also quite an impressive result for 4 million of entries.

5 Results

5.1 Scenarios of use

User: Rachel Dennis

Our user is a young entrepreneur who wants to found a business in the United States of America. Furthermore, the user wants to make sure that the business will receive the maximum number of contracts as well as strong funding from the government. To this end, the user wants to find out what location, time, and type of business is most likely to yield the wanted results.

Using our tool, the user would be able to first filter the charts in the dashboard by selecting the category which receives most funding from the government in the treemap. The bar chart bound to the treemap acts as an additional filter by which the hierarchy can be further explored. Here, single services (products) provided in the selected category can be viewed, compared and selected. Selecting a product would then change the diagrams of the dashboard to only display data from that particular subcategory.

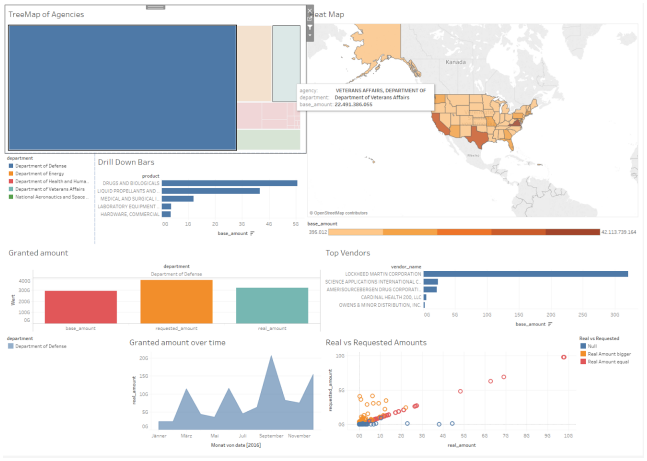

Figure 2: Scenario of use

Using the remaining diagrams, it is now very easy for the user to identify all the information required. Using the heatmap it is easy to identify the perfect location, and using the scatterplot diagram would yield further insight into the type of business most likely to receive government funding. Additionally, the user is able to compare the amount of granted options, further informing the user how generous is the government to which kind of business.

User: Fernando Brock

Our user is the head of an established company doing contract work for the government. When closing contracts with the government a contractor may apply for additional funding in the form of options, which, if granted, would be added to the base amount for the contract. Our user is suspicious that competitors in the same field are getting more options granted for their contracts than his company is.

Using our tool, the user would first filter the displayed data using the treemap to the general category of his company. The user then has the option of filtering down the content to the last filtering level, which is the specific product. At this point, the data changes to display data connecting to the services in the given subset of categories the user has selected.

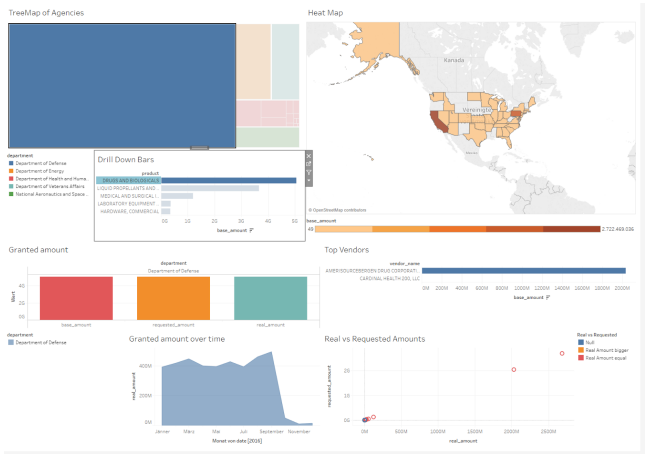

Figure 3: Scenario of use

The user can then see the distribution of grants in the category of his interest in the remaining graphs of our tool. In the heat map the geographical distribution of funding can be examined, while the area chart gives insight into the temporal distribution. Furthermore, the sums of the base, real and requested amount of options for the selected companies are displayed in the bar chart.

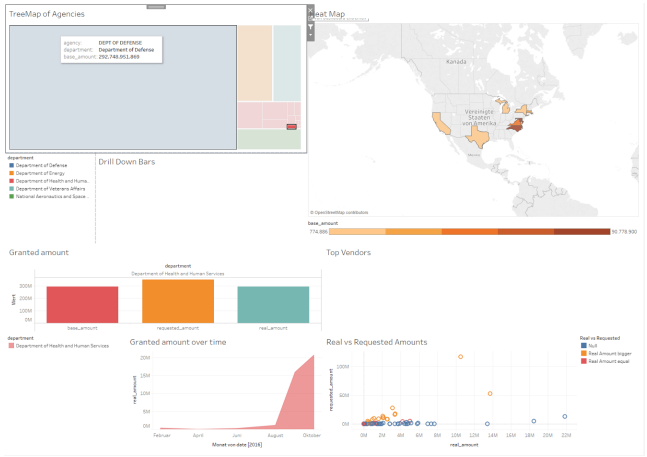

Figure 4: Scenario of use

The user can also use the scatter plot and ranking bar chart to view and compare the top competitors in his category.

User: Antoinette Andrews

Our user is a member of a group representing farmers. Due to the struggling industry in the user's area our user wants to find evidence of the government mishandling funds.

Using our tool, our user would first select the desired state from the heat map, which gives the user an initial impression of how much funding the users state receives, and filters the dashboard. Our tool now displays data connecting to the selected region.

The user can now use the options bar chart and the treemap to search for e.g. an unusual amount of options granted in an area. This could indicate favoritism towards a specific sector and even specific services. The user may also examine the area chart to see how government funding changed over the year, giving further insight.

Furthermore, our user can then identify outliers in the companies in the bar chart and scatterplot displaying company data. This helps

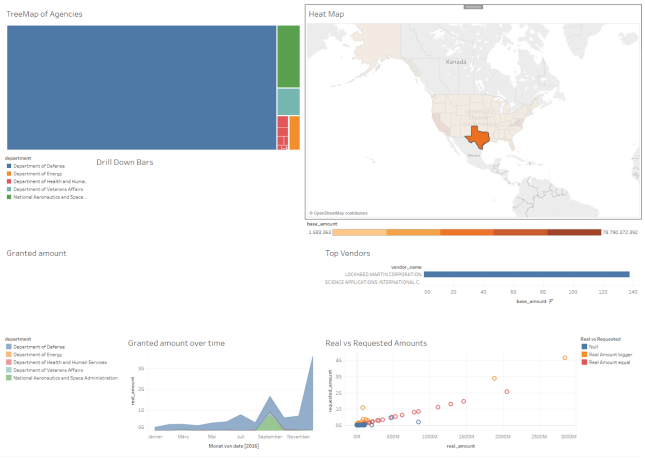

Figure 5: Scenario of use

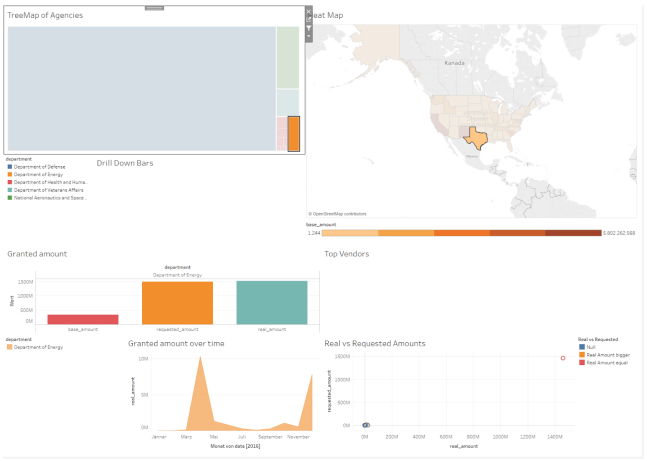

Figure 6: Scenario of use

the user to identify companies which may receive an unfair amount of government funding.

5.2 Performance

As already described, we used the government spending data for our dashboard. The problem with this data set, was that its file size exceeded 10GB, almost reaching 11GB. Due to the sheer number of entries in the unfiltered data, the performance of our dashboard was of concern to us from the beginning. To combat the performance loss caused by such a file size, we decided to filter every field that would not be needed. This left us with only 12 fields, and a file size for our data set of 813MB. Thus, with a file size that could be handled comparatively easy by tableau we measured the overall performance.

Tableau offers a built in tool for measuring performance on a local machine, which we used for our tests. What we discovered during these tests, was that the main part of computation was taking place for the calculation of the scatter plot, taking more than 1 second, while all other computations took less than 100 milliseconds.

This obviously lies in the nature of the scatter plot, since it must display many individual transactions as individual points in the chart.

All in all, we decided that the small performance loss was not worth compromising the scatter plot, since working with the dashboard was not rendered frustrating or even impossible by it.

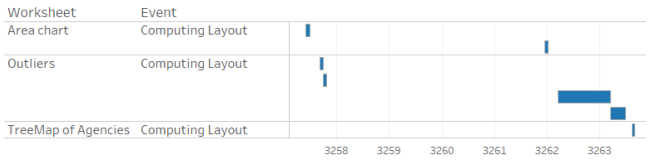

Figure 7: Tableau performance

5.3 Feedback

The feedback we got for our Milestones of course influenced our final design, helping us decide on a specific design and improve functionality overall.

Starting from our first Mockups, we moved away from the Infographic layout, to make space for more detailed views. Furthermore we moved away from the Idea to have what is now a treemap displaying departments being a stacked Bar chart, increasing readability.

Furthermore, based on the feedback we received specifically for the treemap and the implied hierarchy of the data, we worked on a way to display all 4 levels of our data hierarchy. This resulted in the hierarchy being split between the treemap and the Drill Down Bar Chart. Which serves as a display for the lower levels of the hierarchy. Finally, after repeated insistence,we agreed that our names should indeed be contained in the report.

6 Discussion

6.1 Strengths and weaknesses

Through our implementation of the hierarchy through the treemap and the connected bar chart, it is very easy to browse through all the levels of the data, from Agencies down to individual vendors. This also enables easy categorization of the vendors, since they are always shown in context of the selected department. Also, through the nature of our dashboard it is easy to determine the maximum values in a variety of categories, be it temporal or geographical data. Furthermore, the bar chart granted amount and the scatter plot enables the inspection of favoritism on the side of the government. The scatter plot also grants the user the possibility to do rudimentary trend analysis.

Where we find our dashboard lacking lies more in additional functionalities, which would improve the user experience. For example, the performance, while we ultimately decided to keep it as it is(as explained previously), could be better. Additionally we wanted to colour code the map to the general colour code of the tree map, but this cannot be achieved in tableau without considerable difficulties.

6.2 Lessons learned

Throughout the lecture and the corresponding project, we have learned a lot about the design of dashboards, and of the pros and cons of specific graphs or view in specific. Concepts and techniques were introduced which would come in handy in the project, as mentioned above in the report. We have also lost fear of the bar chart, and gained an aversion to the pie chart.

Another lesson came from the manipulation of such a big amount of data. Even though we specialized only on contracts from year 2016, the original dataset contained 11 GB of transactions, which exceeds the capabilities of classical editors. Besides over 4 million of entries, we had to handle 225 attributes of various structure, go through them and decide which are relevant to our use cases. The preprocessing also required finding the hierarchical and other relations among the attributes.

Additionally, we were familiarized with different tools for creating dashboards, and received rudimentary JavaScript knowledge. Though we decided to focus on tableau in our project, we feel we have also gained insight into the d3 platform.

All in all, we learned how to create interesting and informative graphs and how to combine them into a coherent picture.

7 Separation of tasks

Separation of tasks	
Name	Task
Artan Toplanaj	Motivation, Related Work and PDF with Latex
Dominic Palffy	Results and Discussion
Jiri Mauritz	Approach and Implementation

8 Our tool

Our tool can be found here.