CSE 564: Visualization Final Project Proposal

Snigdha Kamal and Prachi Poddar

April 11, 2017

1 Background

DonorsChoose.org is a US based nonprofit organization that allows individuals to donate money directly to public school classroom projects. Public school teachers post classroom project requests on the platform and individuals have the option to donate money directly to fund these projects. The classroom projects range from pencils and books to computers and other expensive equipments for classrooms. In more than 10 years of existence, this platform helped teachers in all US states to post more than 7700,000 classroom project requests and raise more than \$280,000,000. DonorsChoose.org is making the platform data open and available for making discoveries and building applications. In our project, we aim to use this dataset to build interactive data visualization that represents school donations broken down by different attributes.

2 Dataset Specifications

The size of the dataset is 204MB when zipped. It is in the form of comma-separated files which can be read using a normal text editor or Microsoft Excel. The size of the dataset after unzipping is 645 MB. This dataset contains 1,203,287 records and 44 attributes. Some example attributes are:

- projectid: identifies a unique project

- teacheracctid: identifies the teacher who created the project

- schoolid: identifies the school that this project is for

- resource_type: Books, Technology, Supplies, Trips, Visitors, Other

- poverty_level: high, low, minimal, unknown

- grade_level: Grades PreK-2, Grades 3-5, Grades 6-8, Grades 9-12

- total_donations: Total donation amount

- num_donors: Number of unique donors giving to this project

- funding_status: Completed, Expired, Live, or Reallocated

- date_posted: Date a project was approved by staff to be listed on the site

3 Problem Statement

Through this project, we are trying to uncover hidden insights into Americas education system by applying visualization techniques and data analytics to infer trends from the data, spot the outliers and make meaningful sense of the data-points. The aim of this project is to create an interactive data visualization dashboard using the Project dataset from DonorsChoose.org. This data visualization represents school donations across the United States over a certain time period.

The dataset we have chosen has over 7 million rows. Representing data with over a million data points poses a serious problem for real-time visualization. One way to approach this issue is to use subsets of data and an interactive dashboard which provides different visual models on the data and changes the plots in real-time according to any change in the dataset and filters applied. This gives end-user a more flexible and open-ended capability to observe the data, which otherwise seems to be an impossible task. We will be employing random and stratified sampling to condense our data points, making it easier to handle and visualize. We will also draw a comparative analysis between the results obtained from the two sampling techniques.

3.1 Goal of the Project

Our goal is to build a visualization tool that can effectively assist people in observing the significant donations being made in different states of the US and how they have impacted the quality of American Education System over the years.

The following are some of many questions that we will attempt to unearth from our dataset using various visualization techniques:

- Distribution of donations made per year and the amounts donated

- Distribution of donations based on resource - books, technology, supplies etc.

- Distribution of donations by poverty level

- Distribution of donations by state

- Possible linkage of donations to the literacy level of each state

- Total number of donations made and their cumulative amount - per year and per state

Figures 1 and 2 represent examples of dummy plots. We aim to derive similar plots from our dataset and draw meaningful insights.
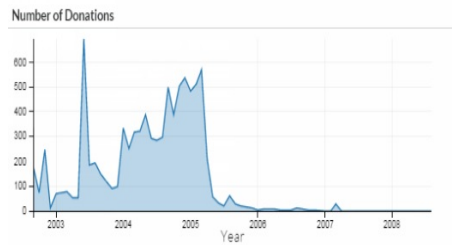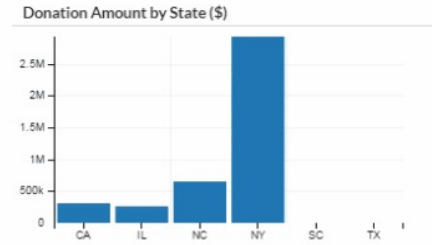
Figure 1: Sample Plot 1

Figure 2: Sample Plot 2

4 Tools and Technologies

- D3.js: A JavaScript based visualization engine, which renders interactive charts and graphs based on the data.

- DC.js: A JavaScript based wrapper library for D3.js, which makes plotting the charts a lot easier.

- Crossfilter.js: A JavaScript based data manipulation library that enables two way data binding.

- MongoDB: NoSQL Database used to convert and present our data in JSON format.

- Flask: A Python based micro-framework used to serve our data from the server to our web based interface.

5 Timeline

Week of Quarter	Milestone
Week 1 (April 9th - April 15th)	Preliminary design
Week 2 (April 16th - April 22nd)	Data Extraction and Visualisation Prototype
Week 3 (April 23rd - April 29th)	Mid-Project Review / Beta Implementation
Week 4 (April 30th - May 6th)	Final Implementation
Week 5 (May 7th - May 12th)	Project Report and Poster Presentation

Figure 3: Preliminary Schedule

5.1 Milestone notes

The preliminary design will have all the design choices made including how to extract the data and which visualization framework to use.

The data extraction prototype will be a reduced quantitative representation of the original dataset present in existing repository. The visualization prototype will be a basic first stab at the core of the interactive dashboard. The two prototypes will show that the design choices are sound and work will continue to the beta stage which will be a working end to end implementation. The last week will conclude with the final cleanup, report writing and poster creation.

References

[1] Parallel Coordinates. URL: https://bl.ocks.org/jasondavies/1341281.

[2] Crossfilter.js. URL: http://square.github.io/crossfilter/.

[3] Dashboard. URL: https://anmolkoul.wordpress.com/2015/06/05/interactive-data-visualization-using-d3-js-dc-js-nodejs-and-mongodb/.

[4] DC.js. URL: http://dc-js.github.io/dc.js/examples/.

[5] MongoDB Documentation. URL: https://github.com/mongodb.

[6] Project Dataset - DonorsChoose.org. URL: https://research.donorschoose.org/t/download-opendata/33.

[7] D3.js examples. URL: https://bl.ocks.org/mbostock.

[8] Flask. URL: https://www.fullstackpython.com/flask.html.

[9] D3.js visualization ideas. URL: http://techslides.com/over-1000-d3-js-examples-and-demos.

[10] Literacy. URL: https://en.wikipedia.org/wiki/List_of_U.S._states_by_educational_attainment.

[11] Chloropleth Map. URL: https://d3-geomap.github.io/map/choropleth/us-states/.

[12] Scatterplot Matrix. URL: https://bl.ocks.org/mbostock/4063663.

Project Progress Report

1 Introduction

Data visualization plays an important role in data analysis and drawing meaningful conclusions. It enables data analysts to effectively discover patterns in large datasets through graphical means, and to represent these findings in a meaningful and effective way. So far, in this project we have implemented the same and more. This project goes beyond pure data analytics and enables the user to interact with the visual plots by filtering thousands of datasets in real time in the browser by refining the dimensions. As discussed earlier, this project is focused on visualizing the Project dataset which contains data points for the classroom projects data available with DonorsChoose.org, a detailed description of which has been given in the subsequent sections.

2 Project Structure and Workflow

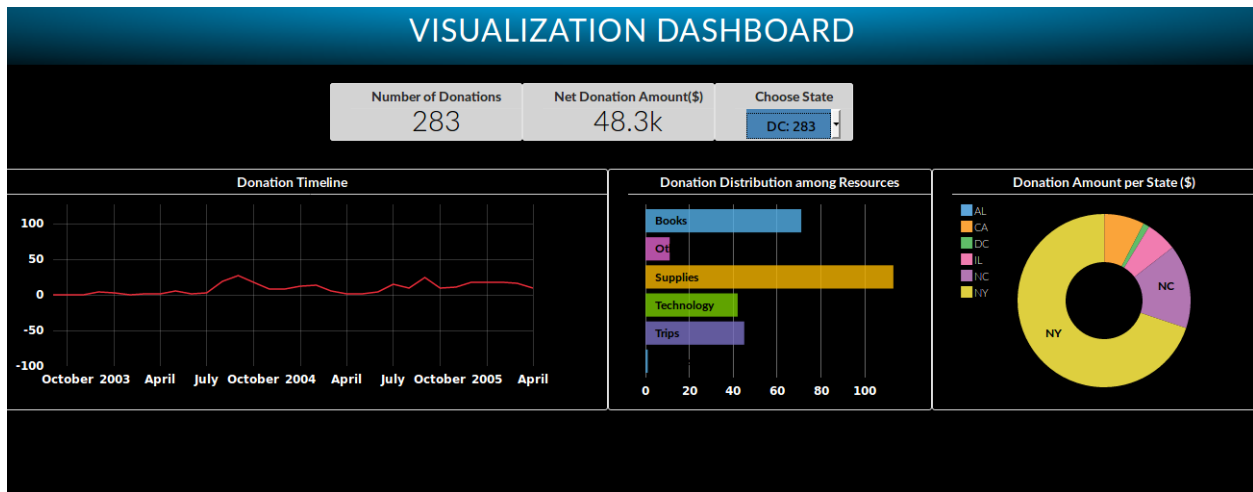Figure 4 shows the dashboard which we have built till now.

Figure 4: DashBoard

2.1 Key Components

The key components of the project are discussed in detail as following: Figure 5

1. The original data is in .csv format. Since it is a large file with over 1.2 million records, only a subset of the original data have been used. The data subset contains of the first 9000 records picked from the main csv file. So far we have used only the following attributes for our analysis:
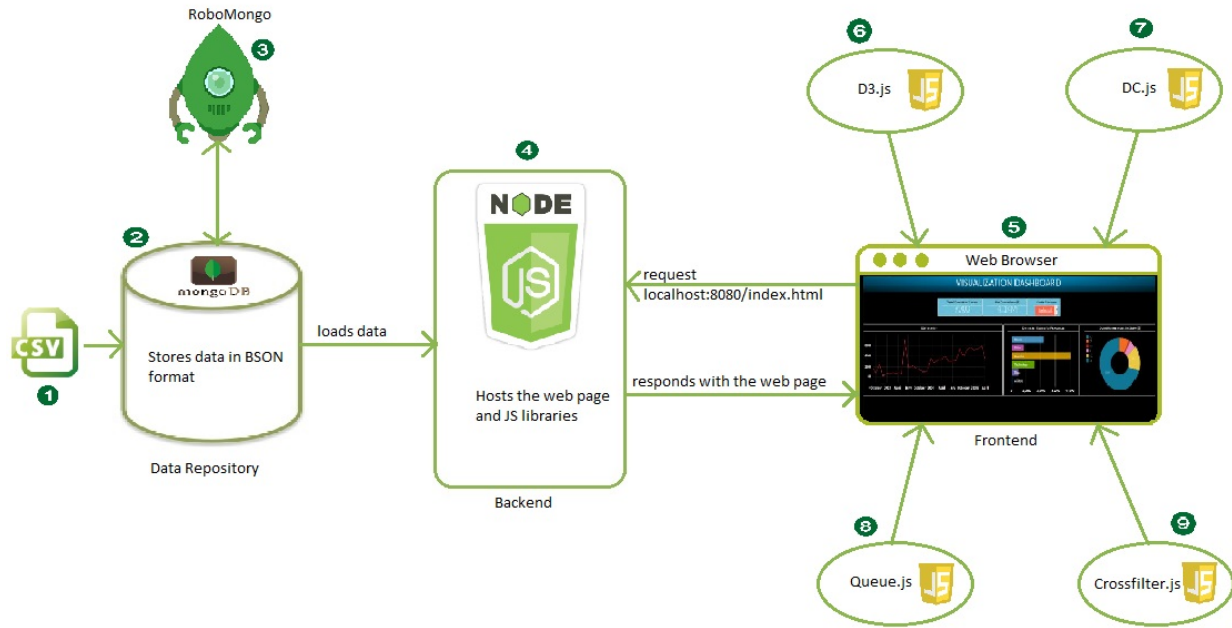
Figure 5: Flow Chart

- total_donations
- date_posted
- school_state

2. MongoDB, one of the most popular non-relational database has been used as our data repository. The reason for using Mongo is its scalability and flexibility in storing records of different formats in the same database. MongoDB stores data in BSON format (Binary JSON), which is a format that is both human readable and easily parsed by the computer. The data is loaded into the database using the following command.Figure 6


```
MongoDB Enterprise > mongoimport -d donorschoose -c projects --type csv --header
line C:\Users\Prachi\visualisation\sample_data.csv
```

Figure 6: Load Data into MongoDB

3. RoboMongo is a GUI based MongoDB management tool. It connects to the running instance of MongoDB using the port and hostname. It helps to browse through thousands of records of data which otherwise is a complicated feat with the command-line. The following figure shows our RoboMongo shell loaded with our dataset: Figure 7

4. The server is built using nodeJS which is one of the best server platforms. The main reason to choose it over other Python/Java based servers is due to its high performance and fast connectivity with MongoDB. It interacts with the database and renders the HTML page that contains the dashboard.
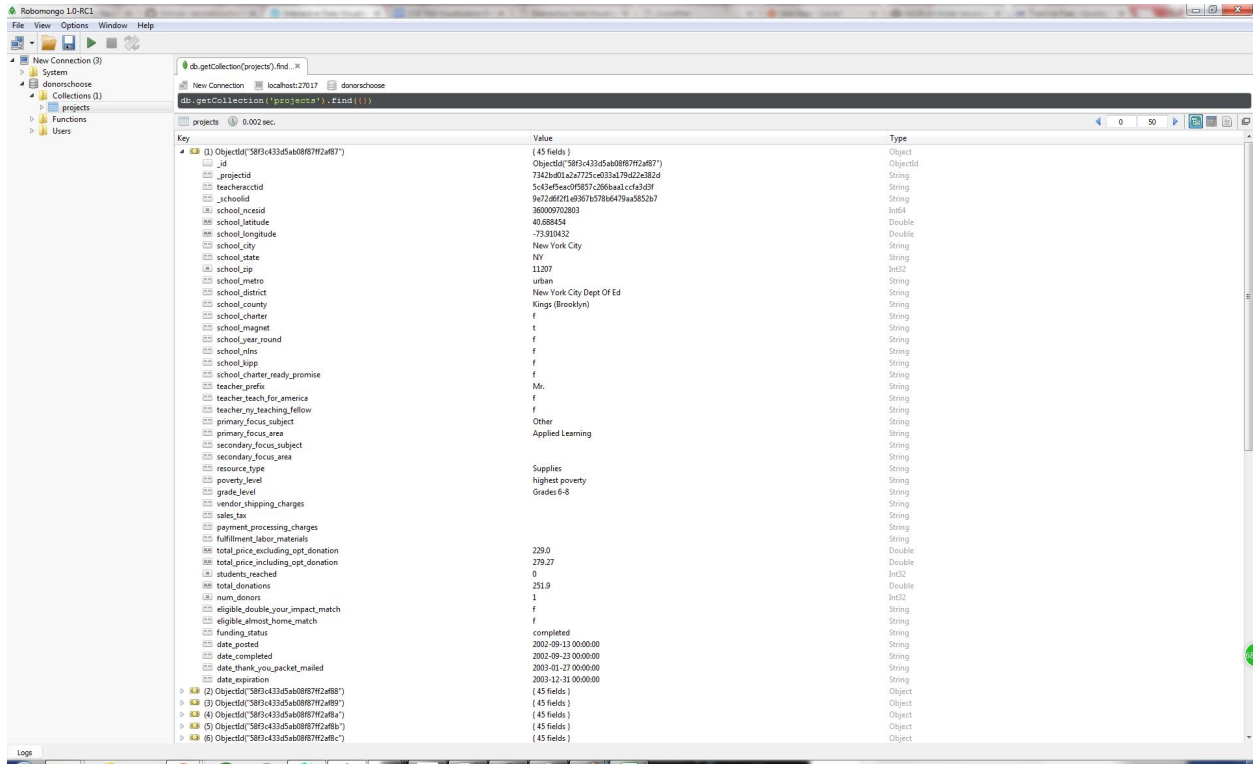
6

Figure 7: RoboMongo Shell

5. The browser on the client side sends the request to the server in the form of a URL and get the HTML page as the response which contains the interactive dashboard.

6. DC.js library has been used to render interactive charts and graphs with great ease. Unlike Processing.js, Paper.js or other SVG-only based libraries, it works seamlessly with existing web technologies, and can manipulate any part of the DOM. A small d3 code-snippet to render total donation count on selected rows is given below: Figure 8

Figure 8: DC.js Code Snippet

7. There are two main reasons for using the DC.js library. Firstly, it is a glue that holds D3.js and Crossfilter.js together. Secondly, it acts as a wrapper for D3.js, and helps in making the code more generic. Plus the DC library has a lot of inbuilt charts to suit majority of analysis. Once all the plots are created, the "dc.renderAll()" function renders all the charts into the web page.

8. Crossfilter.js has been used to make the dashboard interactive and give more control to the end user, thus improving the overall understanding of the data for the user. The

7

data is ingested into a crossfilter instance and dimensions are created based on the instance. It acts as a two way data binding pipeline. Whenever selection is made on the data of specific chart, it is automatically applied to other charts as well enabling the drill down functionality.

3   Progress Summary

The following work has been completed so far:

- Project design: A well-specified design model has been prepared which defines the role of each component and the interconnectivity among the components.

- Data extraction: This included getting the data in csv format, creating a subset, loading the subset into the data repository and finally connecting the database with the web server.

- Project framework: A high-level project directory structure has been created with almost all major components implemented either fully or partially.

- Visualization Plots: Three visualization plots have been completed which are - a line graph, pie chart and a row chart. Along with this, the total count of number of donations and donation amount (in dollars) is also displayed.

4   Interesting Observations

The following are some interesting observations which could be inferred from the data:

- Total donation collected over the years is maximum for the state of New York and least for the state of Alabama.

- Donation Count for Resources
  It gives an idea as to how the donation money is distributed in the purchase of resources. We visualized this distribution for the different states in the dataset and over a timeline and obtained consistent results over all the visualizations. Most amount of money was spent in buying supplies such as stationery, paper etc. followed by technology(desktop PCs, smart boards etc.) and then, books.

- Donation Amount by State
  We visualized the net donations of each state towards their schools over the years. We observed that New York was consistently the highest donator over the years followed by North Carolina. This was observed over the entire time scale of the data chosen. However, New York was the largest contributor (over 96% of the total donations) before the year 2004. Other states started contributing in a significant manner only in the year 2005.

5   Problems Encountered

- Initially, we tried to visualize the entire data but this drastically slowed down the dashboard, and made it unresponsive. Even though all the records were being imported into MongoDB quickly, the loading of data into the web server (nodeJS) from MongoDB was slow. There was a trade-off between handling large data and fast response time. Since the project is more visualization oriented, a subset of the original data (1%) was used which resolved the issue.

- In order to decide between NodeJS and Python Flask as the choice of platform for the web server in this project, we ran a comparative performance analysis. The performance of nodeJS was far better than python flask, both in terms of loading the data from MongoDB and rendering charts on browser. Hence, we chose NodeJS as our server platform.

- Picking the first few records from the entire dataset and performing the analysis doesnt do justice to the entire dataset. We need to come up with a better approach to select data in a way such that doesn't affect the analytical results much, is free of bias and gives a good understanding of the original data.

- The x-axis of donation plot which indicates time, has quarter (one-fourth year) as axes-interval. The axis needs to be made uniform to represent just the years.

6   Overall Assessment of Project Schedule

Overall, the project progress is right on track and in accordance with the estimated schedule. We anticipate that it would be successfully completed in Week 5.Figure 9

Week of Quarter	Milestone
Week 1 (April 9th - April 15th)	Preliminary design
Week 2 (April 16th - April 22nd)	Data Extraction and Visualisation Prototype
Week 3 (April 23rd - April 29th)	Mid-Project Review / Beta Implementation
Week 4 (April 30th - May 6th)	Final Implementation
Week 5 (May 7th - May 12th)	Project Report and Poster Presentation

Figure 9: Schedule

Final Project Report

1   Introduction

In this project, we have built a dashboard that is interactive and user-friendly shown in Figure 10. All graphs are built using the DC.js library with complete cross-filtering and brushing support. All graphs can be brushed and insightful observations can be drawn by brushing and highlighting the desired portion of any graph. A demo of this brushing capability is shown in the video. We have also provided a tooltip facility where the exact value of any quantity in any graph is visible by hovering your mouse over the graph.
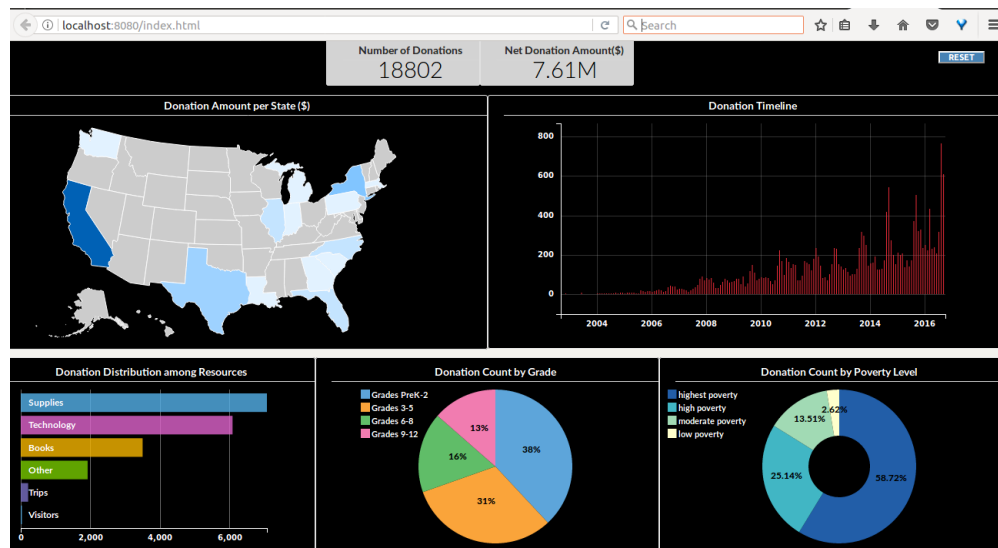
2   Final Dashboard Layout

Figure 10: Final Dashboard

2.1   GeoMap

We used a chloropleth map to represent the donation distribution over each state in the United States shown in Figure 11. The choropleth map provides an easy way to visualize how a measurement varies across a geographic area as it shows the level of variability within a region. A chloropleth map makes for an interactive visualization as it is a thematic map where the areas are shaded in proportion to the statistical variable being represented by the map - the darker the state, the more donations have been made to the schools in that state
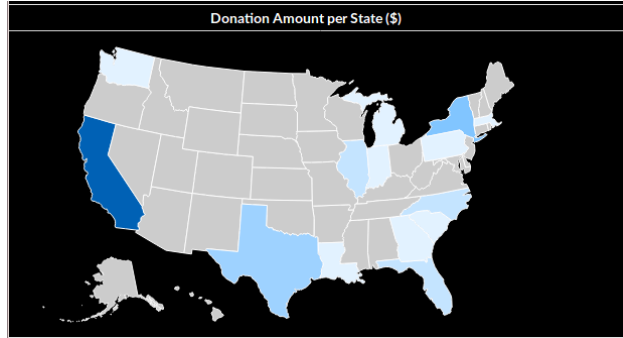
Figure 11: GeoMap

2.2   Scatterplot Matrix

Data comes in a number of different types, which determine what kinds of mapping can be used for them. The most basic distinction is that between continuous (or quantitative) and categorical data, which has a profound impact on the types of visualization that can be used. Categorical/textual data is relatively easy to understand. However, issue comes what there are millions of records having numerical data. For this purpose, we separated all the numerical columns and calculated the eigenvalues to determine which parameters are varying the most. Furthermore, a 3x3 scatterplot matrix for the top-3 highest PCA loaded attributes is plotted as shown in Figure 12. The following are some useful insights which we observed:

1. Most of the high donations are made by donor group size of not more than 2.

2. The increase in the amount of donation does not result in the increase in the total number of children reached. In other words, no matter what the value of donation is, the children reached count is fixed.

2.3   Parallel Coordinates

In order to determine how the numeric data are related to each other, we created a parallel coordinates to plot individual data elements across many dimensions for all 8 numeric attributes. Since relationships between adjacent dimensions are easier to perceive than between non-adjacent dimensions, we have applied the drag and drop feature as well. The data could also be filtered by just selecting a specific range over the axes. The following are some of the useful insights which we observed from Figure 13.

1. All the values in the latitude and longitude axes spans over the geographical coordinates of US. This makes sense as the data is confined only to US. Few outliers indicate remote places like Hawaii, etc.

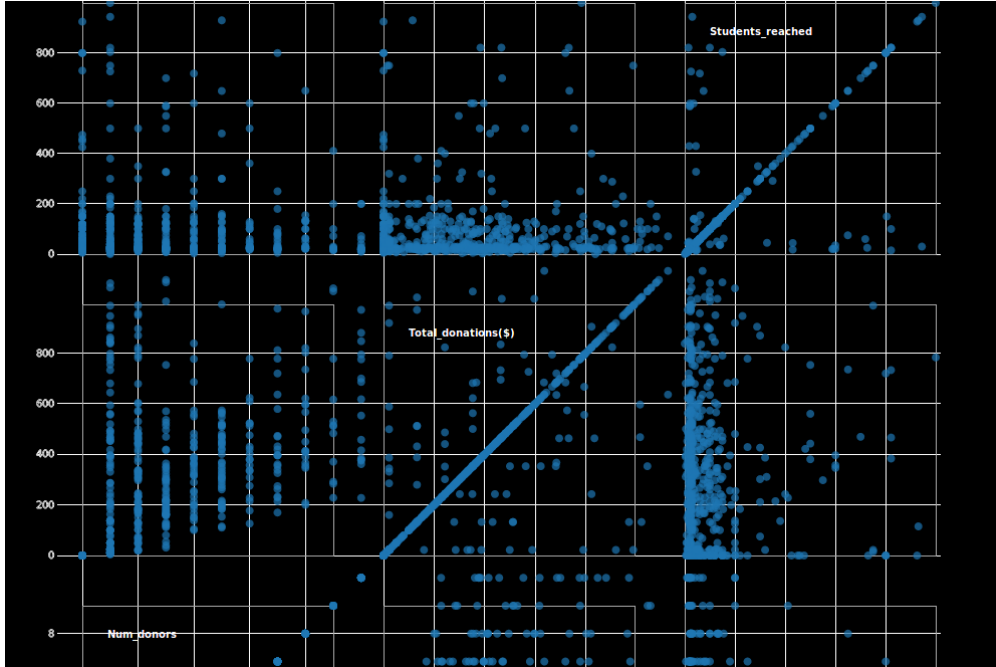2. With a small change in the latitude/longitude, the Zipcode changes drastically.

Figure 12: Scatterplot Matrix

3. There is a strong positive correlation between total price including optional donation and total price excluding optional donation. This clearly indicates that the value of optional donation doesnt vary much across different states.
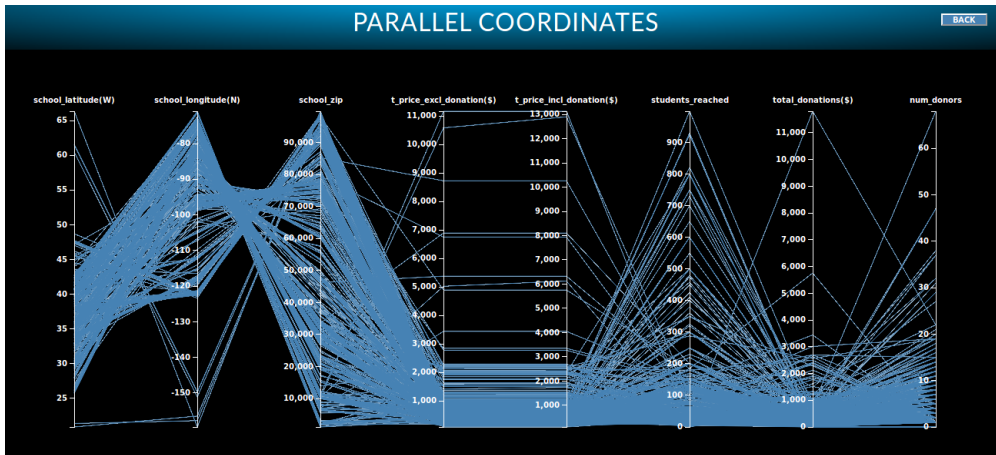

Figure 13: Parallel Coordinates

3   Insightful Observations

The following are some of the observations that we made by visualizing our dataset.

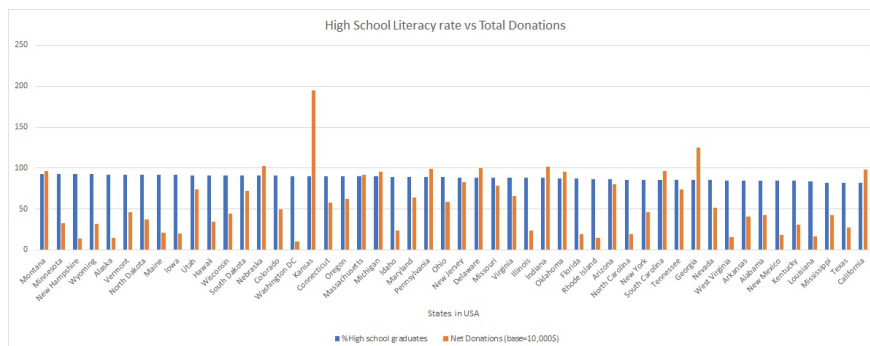## 3.1 Correlation with Literacy Rate per State

Figure 14: High School graduates percentage vs total donations

The main insight that we aimed to unearth from our project was the relationship of the donations in each state to the literacy rate of that state.[10].However, as our data was huge with over 1.2 million attributes, we were able to use only one percent of the data to visualize it on the dashboard without slowing it down. However, we correlated the literacy rate of each state with the net donations in that state separately and the results are shown in Figure 14. It is evident that the percentage of high school graduates stays between 90 and 100 percent for almost every state, irrespective of the donation amount in that state. There is no correlation between the literacy rate and the high school donations in that state. A further research showed that literacy rate of a state depends on many factors such as population density, enrollment rates, average years of schooling of adults etc. and hence, our simple model of high school donations could not describe the literacy rate of that state completely.

3.2 Dashboard Observations

We drew the following observations using our dashboard:

1. Total donation collected over the years is maximum for the state of New York(as is evident by its dark blue color on the map), closely followed by California and least for the state of Alabama.

2. Donation Count for Resources
   It gives an idea as to how the donation money is distributed in the purchase of resources. We visualized this distribution for the different states in the dataset and over a timeline and obtained consistent results over all the visualizations. Most amount of money was spent in buying supplies such as stationery, paper etc. followed by technology(desktop PCs, smart boards etc.) and then, books.

3. Donation Timeline
   Since the inception of this website in 2003, only New York and California were receiving donations. Other states started receiving donations prominently only after the year 2010, indicating that this system of posting requests for donations on the website and receiving money gained popularity only in 2010.

13

4. Donation by Poverty Level
   The schools which posted and received the maximum donations were those with the highest poverty level. This trend was consistent over all the states. This was a good indicator that schools which needed help the most were receiving it in large numbers.

5. Donation Count by Grade
   This pie chart shows that most of the donation resources were being used for Grades PreK-2, consistently over all the states.

4   Conclusion

In this project, we successfully created a dashboard and drew several important observations. We learned how a good and efficient visualization can help bring out crucial relationships which otherwise remain hidden from the human eye. An interactive visualization makes it easy for anyone with no prior knowledge to understand and play around with the data.

References

[1]   Parallel Coordinates. URL: https://bl.ocks.org/jasondavies/1341281.

[2]   Crossfilter.js. URL: http://square.github.io/crossfilter/.

[3]   Dashboard. URL: https://anmolkoul.wordpress.com/2015/06/05/interactive-data-visualization-using-d3-js-dc-js-nodejs-and-mongodb/.

[4]   DC.js. URL: http://dc-js.github.io/dc.js/examples/.

[5]   MongoDB Documentation. URL: https://github.com/mongodb.

[6]   Project Dataset - DonorsChoose.org. URL: https://research.donorschoose.org/t/download-opendata/33.

[7]   D3.js examples. URL: https://bl.ocks.org/mbostock.

[8]   Flask. URL: https://www.fullstackpython.com/flask.html.

[9]   D3.js visualization ideas. URL: http://techslides.com/over-1000-d3-js-examples-and-demos.

[10] Literacy. URL: https://en.wikipedia.org/wiki/List_of_U.S._states_by_educational_attainment.

[11] Chloropleth Map. URL: https://d3-geomap.github.io/map/choropleth/us-states/.

[12] Scatterplot Matrix. URL: https://bl.ocks.org/mbostock/4063663.