Interactive Systems 3

Data Analysis Tool Report

5 Dec 2013

Team members:

Martynas Buivys
Najam Hussain
Mircea Iordache
Vlad Schnakovszki
Robert Szkokan

Table of contents

1. Design Requirements

The software system will be a data analysis tool. The system will contain a data loading system allowing data to be selected and refined, graphical displays of the data using a scatterplot and the current state of the data being analysed and automated systems allowing correlations to be found. The users will be using the system to find out more about the WHO and Olympic data.

The intention was to develop a simple but powerful tool to help in determining correlation between two sets of data. We decided to take as a previous example the TradeViewer application we were given to enhance in Second year.

Thus, the system has two main components: the data visualisation framework and the refinement tools that allow the user to perform various operations in order to explore in depth the system.

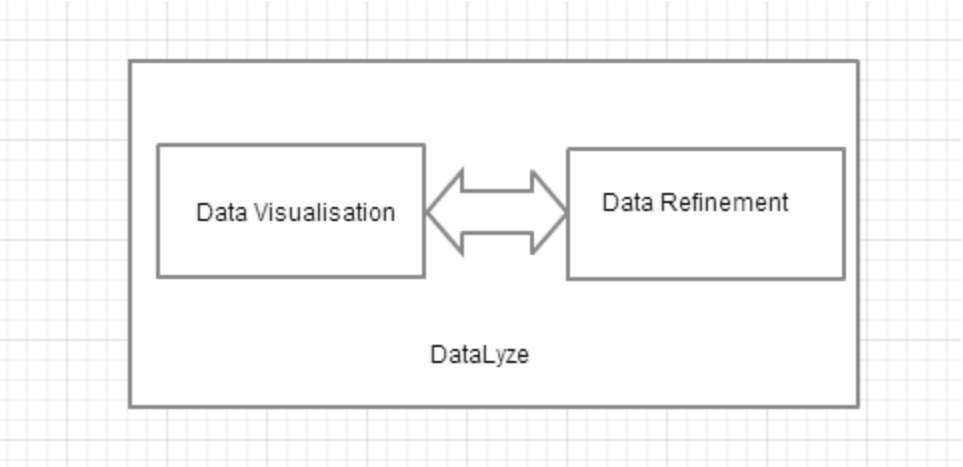

Fig. 1 The System

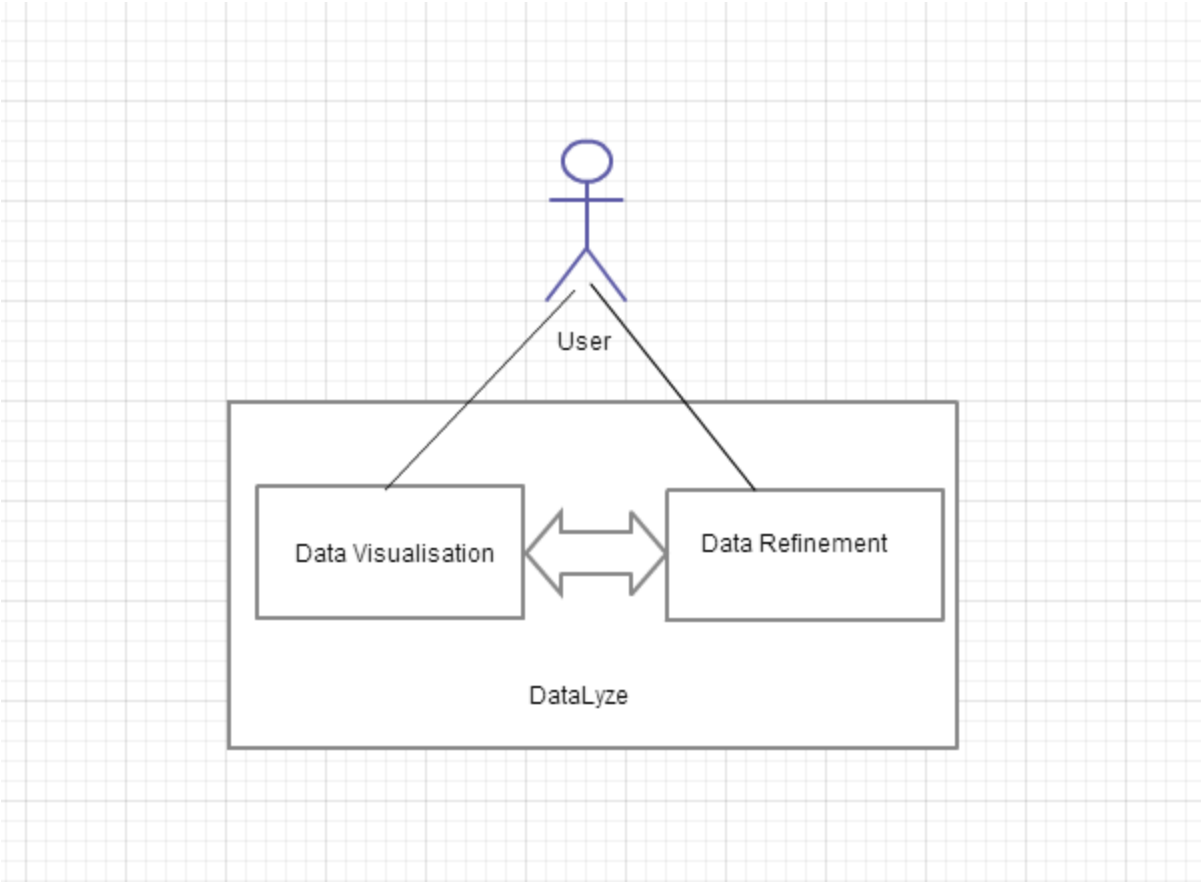

Fig. 2 The Environment

In order to visualise the data, we decided that our main component to do so is a scatterplot, which should able to pan and zoom to show multiple values or the exact position of a tight cluster of values. This, although simplistic approach, provides a good opportunity to provide insight on the data given.

In order to visualise the data in a meaningful manner, we had to implement a way to select pairs of datasets. Another feature we wanted to have is the calculation of the correlation coefficient, which shows how strongly linked are the two sets of data provided.

In order to do in-depth analysis, by removing irrelevant values, a filtering method had to be provided. We have thus decided on implementing data range filters, to select only data that has values only between certain bounds.

Another addition in regards to filtering is the use of location-based filtering, to be able to further analyse the data by comparing and contrasting. This is divided by regions, such as Europe, Africa, etc. but also by the ability to select different countries that can be

shown in the system. The main advantage of this is to better understand the position of a certain country in relation to others, or to 'benchmark' a country, compared to a user-defined 'normal value'.

The combination of the above techniques is already starting to become a template for a powerful and effective data analysis tool.

In terms of non-functional requirements, we have decided upon the following. As the software is not resource intensive, performance is not a concern. The software should behave normally and consistently. Errors should be shown to the user. For the purposes of this project, security is not a concern of ours. The software will be completely available to any machine with an internet browser and internet connection.

The final overview of the system was created, based on the requirements shown, in the image below.

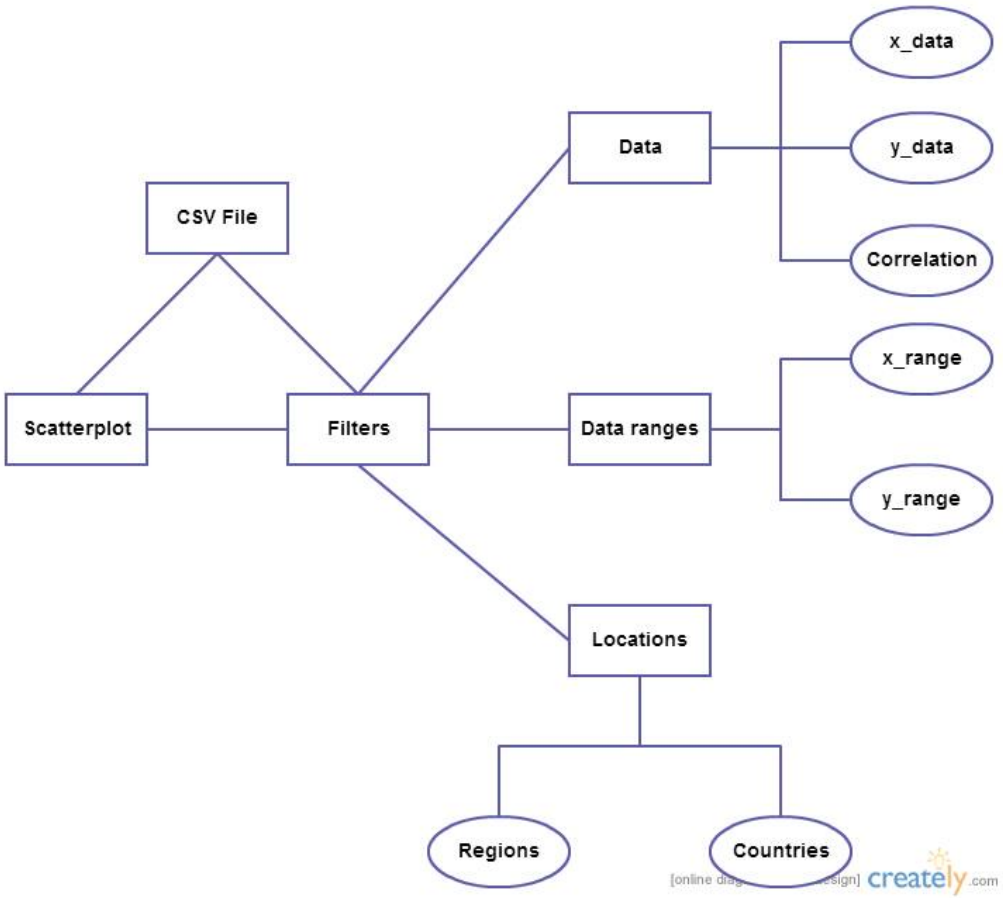

Fig. 3 Top-Level Design of the system

2. Description of the System

With the requirements above stated, we decided on making a web-based system that allows users to interact with the system.

The software is designed to take no input file from the user, as the data is stored within the system, in order to prevent errors to appear. The only interactions that the user is allowed to do are related to the exploration of the data and finding out details that are given.

The system is designed as a web page, with underlying JavaScript files that allow the visualisation and refinement.

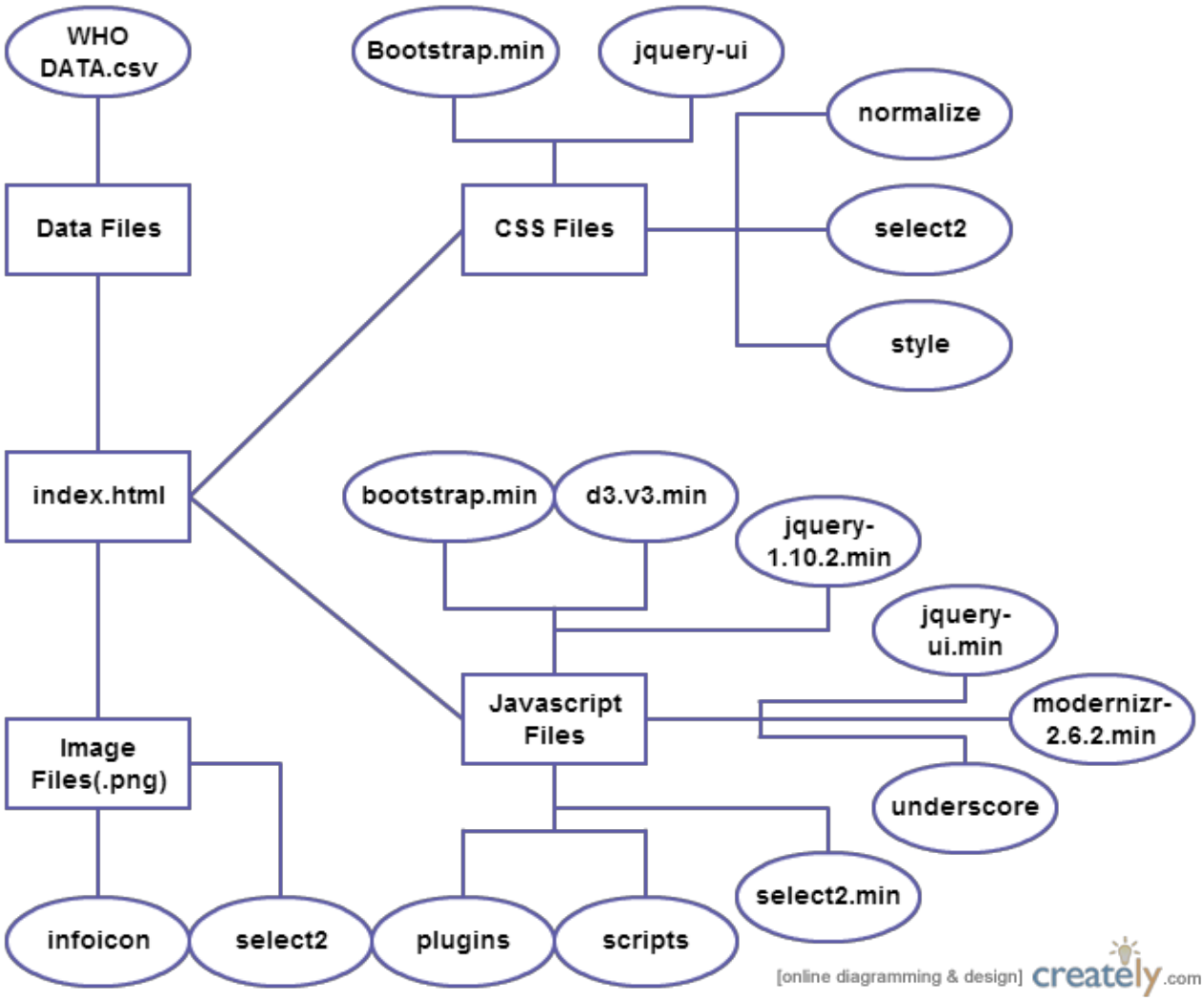

Fig. 4 Dependencies required to run the system

index.html

This is the wrapper that holds all the components. It also uses CSS files for formatting, which are irrelevant to the functionality of the system.

main.js

Creates the scatterplot by loading the data from the server, selecting the datasets, setting the graph size, finding out the minimum and maximum values for the selected datasets, updating the range slider used for data refinement to have the previously-found datasets minimum and maximum values, defining the scale, and finally creating the graph. Once the graph is created, the zoom() function handles panning and zooming when the user interacts with the system.

scripts.js

Handles with location-based filtering and updating the scatterplot after the selections are made.

select2.js

Initializes jQuery UI components and provides functionality to all elements in the system.

underscore.js

An additional javascript library that provides various helper methods.

We have also made use of the D3.js framework and jQuery to integrate JavaScript and HTML so that our system works seamlessly.

3. Analytic findings

Correlation coefficient	Variable 1	Variable 2	Conclusions
+0.74	Agriculture contribution to economy	Children per woman	The more a country relies on agriculture as its financial source, the more likely a woman is to have many children. This might be because agricultural work requires unskilled workers and children could easily help the family.
-0.64	Agriculture contribution to economy	Net primary school enrollment ratio (female)	Children have to sacrifice their primary school education in countries based on agriculture.
-0.60	Agriculture contribution to economy	Net primary school enrollment ratio (male)	Children have to sacrifice their primary school education in countries based on agriculture.
-0.70	Agriculture contribution to economy	Fixed line and mobile phone subscribers	Countries that rely heavily on agriculture use less phones.
-0.70	Agriculture contribution to economy	Healthy life expectancy (HALE) at birth (years)	Children born in agricultural countries tend to have a lower life expectancy.
-0.70	Agriculture contribution to economy	Income per person	People living in agricultural countries are poorer.
-0.62	Agriculture contribution to economy	Literacy rate youth total	Literacy rate is lower in countries that rely on agriculture. This may be because children are guided to work rather than study.

-0.63	Agriculture contribution to economy	Population in urban areas	The population tends to live mostly in rural areas in countries with a large agriculture contribution. This might be because agriculture is prevalent in rural areas.
-0.68	Agriculture contribution to economy	Population with sustainable access to drinking water	Countries that rely on agriculture as a large source of income have poorer access to drinking water.
-0.73	Agriculture contribution to economy	Sugar per person	The more a country relies on agriculture, the less its citizens' sugar intake is.
+0.69	Agriculture contribution to economy	Years of life lost to communicable diseases	Citizens of countries that rely heavily on agriculture are more likely to die because of communicable diseases.
-0.68	Agriculture contribution to economy	Years of life lost to non-communicable diseases	Citizens of countries that rely heavily on agriculture are less likely to die because of non-communicable diseases.
+0.81	Arms imports	Coal production	Arms imports are more significant in countries that have a large production of coal.
+0.60	Arms imports	External debt total DOD current (USD)	Countries that have a larger external debt tend to have more arms imports.
+0.91	CO2 emissions	Energy use	As expected, countries that have larger CO2 emissions also use more energy.

+0.73	CO2 emissions	Income per person	Countries with higher income tend to produce more CO2.
-0.79	Children per woman	Adult literacy rate	The more children women have, the less likely it is that they will be literate.
-0.79	Children per woman	Literacy rate youth	The more children women have, the less likely all of them are able to receive proper education.
-0.80	Children per woman	Births attended by skilled health personnel	In countries that have more children per woman, skilled personnel are less likely to be able to attend a sufficient number of births.
-0.81	Children per woman	Contraceptive use	As expected, the use of contraceptive reduces the number of children a woman has.
-0.85	Children per woman	Healthy life expectancy (HALE) at birth (both sexes)	The more children a woman has, the less likely they are to live a long life. This can be attributed to the fact that the parental and financial attention is less focused on each child.
+0.63	Democracy score	Math achievement 4th grade	Democratic Countries have better educational systems.
+0.74	Forest area	Number of environment and public health workers	Countries with a higher forest area tend to have more people that work on preserving the environment.
+0.70	Urban population	GDP2011	Countries with a high urban population have a

			higher GDP.
+0.86	Urban population	Total CO2 emissions	The urban population uses CO2-producing technologies at a higher rate.
+0.86	Primary completion rate total	Life expectancy at birth	Countries with improved educational systems tend to have citizens with a higher life expectancy.
+0.74	Primary completion rate total	Healthy life expectancy (HALE) at birth	Children which complete primary education are more likely to live a healthy life.
+0.75	Primary completion rate total	Contraceptive prevalence	Persons with primary education are more likely to use contraceptives.
+0.75	Primary completion rate total	Children per woman	Children that have less siblings are more likely to complete primary education.
-0.63	Primary completion rate total	Agriculture contribution to economy	Children have to sacrifice their primary school education in agricultural countries.

4. Introduction to the System

Our system is designed with usability in mind. It comes with a simple User Interface that shows the Scatter Plot of the Data and Analysis tools.

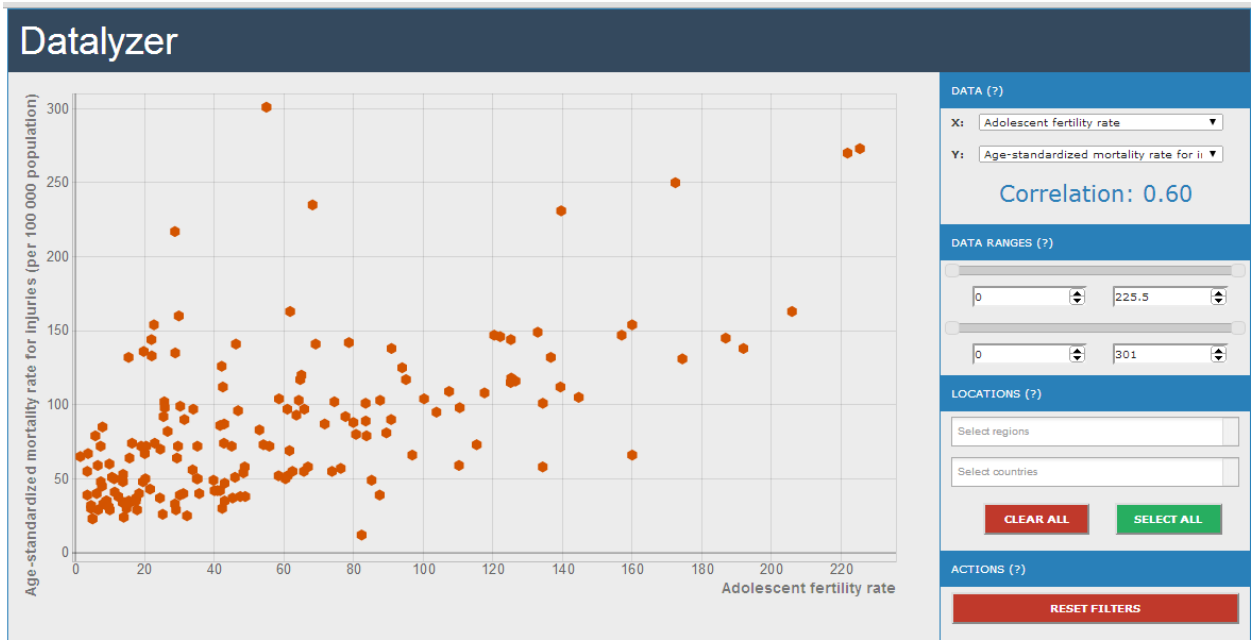

Fig. 5 The system

The Scatter Plot itself allows moving by using the click-and-drag method, and zoom actions using the mouse wheel, or alternatively by double-clicking.

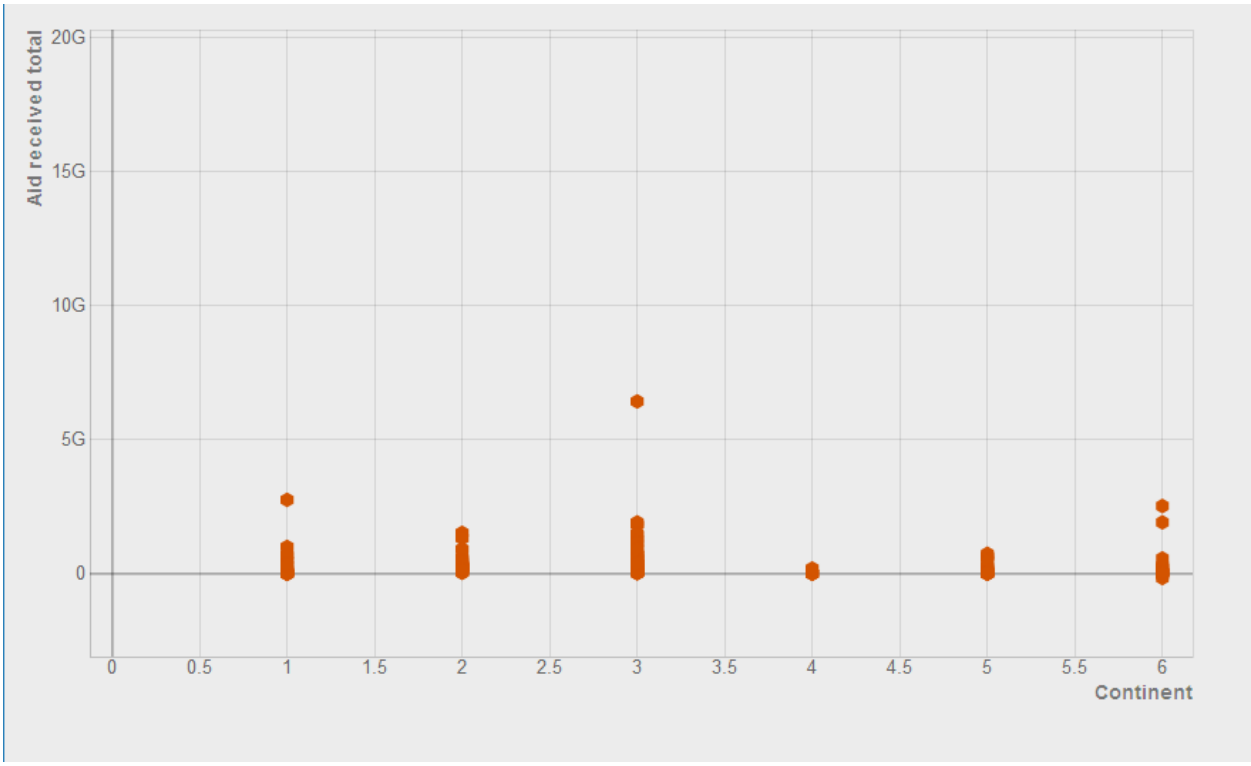

Fig. 6 The Scatter Plot

The analysis tools allow the user to drill down or roll up the data.This is achieved by using the X and Y drop-down boxes to select the sets of data that the user can inspect and see if there is a correlation. This is simplified by the system to show a basic Correlation Coefficient field if a correlation is found.

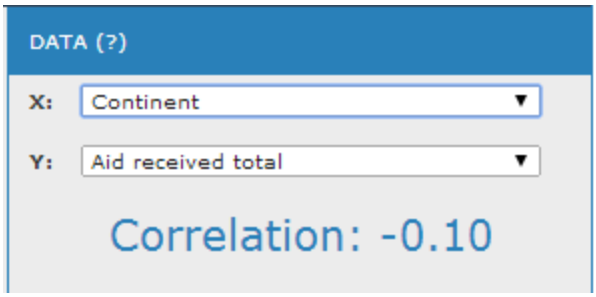

Fig. 7 Data Set Selection and Correlation Coefficient

As well, the user is allowed to set the boundaries for the values using the sliders in the Data Ranges tabs. Alternatively, for precision, there are text boxes so the user can select exact values for the data ranges.

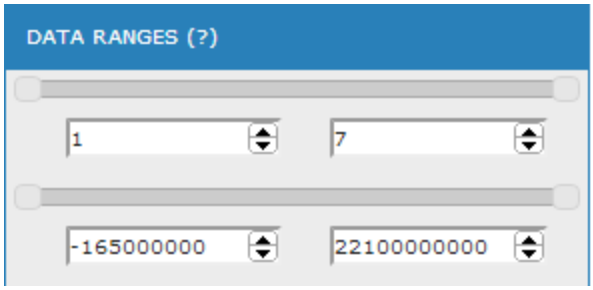

Fig. 8 Data Range Refinement

To inspect values for certain regions or countries only, the drop-down boxes found under the Locations section allow them to select multiple countries or regions, as well as selecting everything or nothing, by using the Select All or Clear All buttons.

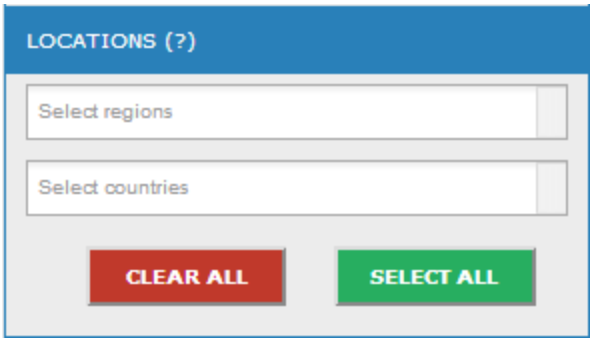

Fig. 9 Location Refinement

In the event that the user wants to reset the tool to its initial state, by clearing out all the data range filters and location-based selections, the Reset Filters option will do that automatically.

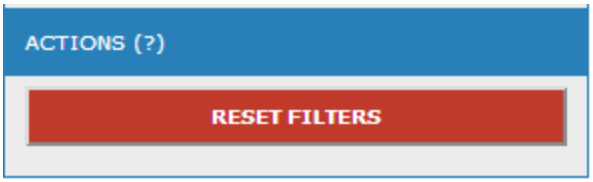

Fig. 10 Filter Reset

5. Acknowledgements

For the general interaction handling and integration of JavaScript and HTML, we have used jQuery and jQuery UI.
http://jquery.com/
http://jqueryui.com/

We used the code provided in the link below as a starting point for our graph implementation. We modified the code and adapted it to our needs.
http://bl.ocks.org/richardwestenra/129f64bfa2b0d48d27c9

For the drop-down checkboxes found in the Location Filters, we used the select2 library.
http://ivaynberg.github.io/select2/

For the feature detection we have used the Modernizr framework.
http://modernizr.com/

6. Team Contributions

Martynas Buivys and Robert Szkokan built the system as they were the only ones familiar with JavaScript and complementary technologies (HTML, CSS, JQuery).

Vlad Schnakovszki did the data analysis and helped with all the other parts required in developing the application and the presentation.

Mircea Iordache wrote the report and the presentation.

Najam Hussain only had minimal contribution in the prototype evaluation.