***GEOG384 Assignment 5 Report***

***Group 1: J Garrah, Sophie, Deboleena***

*Describe the decisions you made in choosing your topic and visualization.*

The data that was obtained for this project was sourced from Environment Canada’s database in their open government web platform. This data was chosen due its format and resulting convenience, its potential for visual display and its relevance in terms of being able to connect it to an enticing story. More effective use of data visualization is done through emphasizing certain trends, which can be deciphered from the different properties available. This is why the pollution incident data was chosen - there were several key sets of information that our visualisation was able to capitalize - the magnitude of the event, the contributing sectors to water pollution and the regions. Even before creating this visualisation, a potential narrative was created focusing on the most polluting areas, the largest polluters and the sectors that contributed the most incidents in 2015. This data was accessible in multiple useful formats for geographic analysis including CSV and JSON - both of which could be queried.

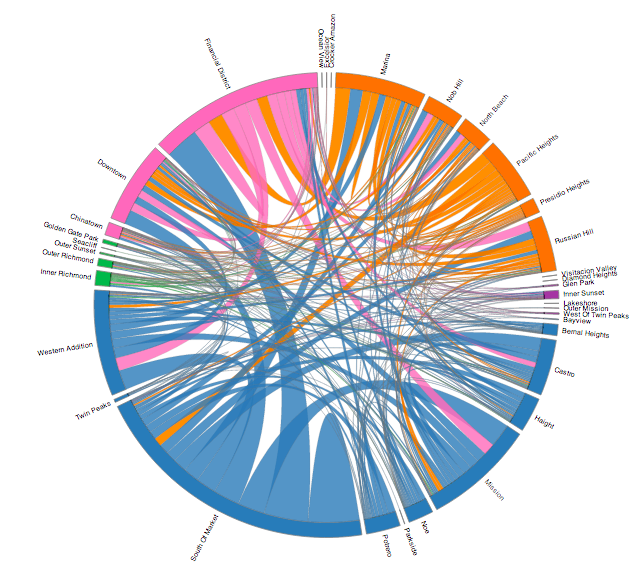
The choice to write a story concerning pollution incidents was also purposeful because environmental data sets are often accessible through government websites and the push for open data regarding climate change is a strong and important movement - especially considering the efforts to save the data recently removed from the American government website. In Canada there is a large amount of published datasets from the environment Canada portal, however for the average person, these datasets are either out of mind, incomprehensible or will not reveal significant facts. Writing an article about the datasets is a powerful way of connecting people to the scientific community and increasing people’s engagement in climate action.

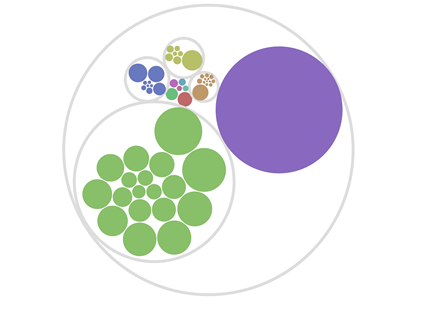
In terms of choosing the actual visualisation type, a bubble chart seemed to be most suitable. The dataset used was incidental and independent so there was no need to create a hierarchy or connections between the nodes. Furthermore, the use of the size of the bubbles for a representation of incident magnitude seemed intuitive due to elegance and simplicity - yet powerful due to ability to express a narrative clearly. The bubble chart allowed us to use both colour and size to express the data narrative whereas the region sorting allowed us to incorporate spatial distribution as a third dimension of visual representation. Lastly, the bubble chart allowed for easy interactivity through mouseovers that cause information boxes to be displayed with the relevant information.

*Report on up to three different visualizations you considered before you chose the visualization you did.*

We considered using a chord diagram to link major polluting industries/entities to the corresponding region in which the incident occurred, but this did not do justice to our theme which was to emphasize the size and range of different kinds of pollution events. Furthermore, since some industries are heavily concentrated in certain provinces and certain provinces have much smaller populations and area, this unequal distribution would cause large gaps in the visualization with few connecting strands. These gaps may bias a viewer to view those smaller provinces as exceptionally well managed or regulated, although this might not reflect the reality of pollution in the water bodies of that province.

Another visualization which we studied in our research for the project was the data packer model. The ability to represent individual incidents as circles appealed to us and this was a feature we would go on to choose in the visualization we finally developed. However, the data packer model has an in-built hierarchy, wherein each circle is positioned within a larger circle. We considered using the biggest circle as a national boundary for Canada and the systematically smaller circles as provinces and eventually pollution incidents, but this approach was heavily focused on spatial distribution as would be a cartographic model. Since we wanted to focus more on the nature and relative size of the spills, we elected to choose another visualization that would better represent the data.

*A sample chord diagram*

*A sample circle packing visualization*

*Defend your visualization choice. Argue why your choice may be superior to a cartographic approach.*

This visualization ended up providing deeper insights into the dataset that we did not predict and were not able to recognize from looking at the data itself. We chose this visualization because we predicted that the combination between categorizing sector (colour), magnitude (bubble size) and geographic distribution (position according to region) would show us the biggest polluting sectors in each region, for example oil in Alberta or pulp and paper in Quebec. Interestingly, we found that in every region the largest water pollution events were caused by municipal governments and wastewater/sewage systems. Where we expected to find differences by region, there were a lot of surprising similarities. This approach was superior to a cartographic approach because it was *not* the spatial distribution that was most interesting but rather the difference in magnitude by sectors (with wastewater/sewage being the largest). The visualization brought the dataset “to life” and the interactivity of splitting by region enhanced the reader experience further by demonstrating that similar patterns hold in every region across Canada. These are insights that anyone would be unable to see from the data alone or even by looking at a GIS-style map of the point data (even if it was colour coded).

***Individual Reflections***

***J Garrah***

This assignment presented us with the challenging task of not only rethinking our conception of “mapping data”, but also with learning and understanding a new framework for displaying data on the web. In the past 4 assignments, when working with data in a cartographic way, all the data analysis and visualization was carried out in Javascript in a mapping API -- it all lived on the map div and was manipulated with various library commands. Now, with d3.js, the data was visualized via an SVG element in the DOM and the d3 library was used to connect the Javascript and the DOM for effective data visualization. I think this was the trickiest part -- learning about the technical side of programming between the Javascript and the DOM rather than containing all the data in a map div and API.

I think one nice aspect if this assignment was to be redone is to change the interactive ‘split’ of the bubbles to not be by region but rather by province. We did attempt this, but dividing the screen into 13 sections (one for every province or territory) was quite difficult, especially when the vast majority of pollution events took place in larger provinces (ON, QC, AB, BC) with smaller provinces only having a couple events (or none at all!). I think dividing by region still shows some interesting trends and was technically much easier to do. Another improvement that might be made is an attempt at data standardization (perhaps per capita?). It’s clear that the Eastern region has much less pollution events but also much less population so it would be a major improvement to standardize the values for good comparison. However, for this project we used the data and numbers provided in the dataset obtained from Environment Canada so that is sufficient for now. To standardize, we could do manipulation right in the spreadsheet before exporting as a .csv or include the analysis right in the Javascript (wouldn’t be too hard at all).

D3.js and the visualization constructed proved to be a powerful way to analyze this data -- insights were gained (such as the fact that the largest pollution events came from municipal governments and wastewater plants) that would be tough to recognize if just spatial distribution was analyzed on a map. Mapping (of course) has the strength of placing these events in our mind in terms of space but the strength of d3.js is that every aspect of the visualization is there to tell the story of the data and the interactivity is designed to enhance this story and make things more clear to the reader (whereas maps require a little more interpretation). It was interesting to learn about a different way to ‘map’ data that is still geographic and I think both of these techniques can be used to complement each other to tell important stories and create new insights into data.

**Sophie Toor, 260612725  
Individual Reflection**

Personally, I found that the most difficult part in the project was understanding the syntax of D3JS and the process of creating an SVG using DOM documentation. As a relatively novice coder, translating the concepts that I had gained an understanding of through java script to d3 notation was a difficult mental process. For example, when creating a constructor type segment of the nodes, I was very comfortable doing so to create a simple object and accessing the properties, but it was a big challenge for me realise how this constructor format firstly functioned in d3, and secondly how it differed syntactically. Going through different tutorials, on youTube mainly, and going through the commented source code examples on GitHub made this process slightly easier and more feasible in my mind. It was also difficult deciding on which type of chart best represented the chosen data narrative.

The functionality in our code that splits the bubbles into groups based on national regions is a very nice use of spatial configuration in data visualisation however I believe that another additional functionality would be having the bubbles split into different groups based on their colour, which is representative of their sector. This additional function would add different ways for readers to visualize the data and deepen comprehension of the story. This allows the reader to understand both the most polluting region and industry merely through interaction with the visualisation. Another improvement that could be made is through an increase in the information provided. It would be more useful to readers if there was information about the actual event provided such as a link to the company’s website and a short blurb about the polluting incident. However, the dataset was limiting and it would be time consuming to provide a url to each company. Finally, another improvement may be to include a pie chart that shows provincial contribution and sector contribution because seeing the data in proportional terms is useful for making broader generalizations in the story.

In terms of data visualization of spatial information, most people jump immediately to the power of cartographic displays. Creating maps is extremely useful for displaying information in a manner than highlights spatial relationships and can use various tools such as colour (of classes for example) and shape (of icons for example) to highlight certain trends. There are multiple benefits such as providing a recognizable and familiar framework for users to see the data. However, having a map requires the data analyst to conform to a certain method of visualization, bound by the projection and the relevant boundaries. This constrains the analyst in the type of narrative they want to emphasize. This is why D3 is useful, there is a lot less limitations in terms of the way different stories the data tells can be emphasized. For example, as previously mentioned, the functionality of the bubble chart to display the magnitude of the event is easily highlighted in D3JS visualizations, but this may be difficult on a map - for example if two incidents were close together and different icon sizes illustrated the magnitude, the points may overlap on a map - reducing the power of the visualisation. D3JS is also highly customizable depending on the data and depending on the chart type used, different aspect can be emphasized. For example the circle packer and zoomable sunburst charts can emphasise a hierarchy, which is sometimes difficult to do using a map. However, D3JS it is a disadvantage that spatial organization of D3JS data doesn’t conform to the recognizable spatial relationships that you can see in a map. Furthermore, it is a very difficult library with a steep learning curve - although there is open source code, it requires a certain level of technical knowledge.

***Deboleena Mazumdar***

The most difficult parts of the project were familiarizing myself with the new vocabulary and methods of D3JS and trying to relate our pollution data to the code written by developers who created the visualizations for their own purposes. I found that functions which I knew how to do in javascript (going through an array of different attributes to select one using a for loop) were done differently in D3JS and although the lines of code may have been fewer and more efficient, it was still a learning curve. Another challenge for me was identifying which parameters we really wanted to display from the large dataset and which would be the best way to do so. For example, we debated whether we should color code by province or type of polluting incident, and decided that type of polluting incident would be more interesting since data is usually grouped geographically in maps. This also allowed for a clear representation of trends in the most culpable entities (private or public). If I were to suggest an improvement for the assignment, it would be that we could explore platforms outside of D3JS such as Tableau to see how other environments compare for creating compelling visualizations from the same data. While I understand the choice of D3JS because of the wide range of functions that it can be used for, as well as a reliance on the kind of coding skills we have seen in the course, it might be interesting to see firsthand what the limitations and strengths of other programs may be. The strengths of working in D3JS and other visualization environments compared to traditional cartographic models is that people are used to seeing maps, and they therefore tend to gloss over them. With visualizations, however, the creator can choose which elements are most pertinent (not always geography, as with our project), and this counter-intuitive presentation of the information can intrigue the viewer into seeing a given phenomena in a new way. This may prompt them to interact with the visualization to reveal more information (which is why dynamic and interactive visualizations are employed, depending on the format and content). The strengths of maps are that they make spatial trends like clustering or equal distribution clearly visible. Moreover, maps can be easily supported on most devices whereas it can be trickier to represent a visualization well on small mobile devices.