Akash Kumar

November 9th, 2017

DataVis. Stage 2. Grafiti Draft

For this project I chose to work with Grafiti, the social media app for data visualization. Their mobile first approach and use of a vast variety of data sets intrigued me and that is why I chose to work with them. A few other people chose to work with Grafiti as well. In order to give them the best product possible and use our limited time with them effectively we all decided to team up and tackle a large data set for them. Together we can create a much more well-rounded solution that will serve Grafiti better than anything one of us could create alone.

As a result of this partnership we were able to tackle the problem differently. Grafiti is working on creating a system of data sets and visualizations that can allow everyday people to use data, rather than emotion, to tell their stories on social media. By creating an app, they put themselves inside of a system that allows for sharing, and makes it easy for people to contest fake news with real data sets. The app itself is also designed to educate the user on the pitfalls of trusting a data visualization. By letting users easily and quickly manipulate data sets they learn how the same data can tell different stories based on how it is visualized, and what key elements are implemented and which ones are hidden. To them data was previously much harder to get to and certainly harder to visualize, they are lowering the barrier significantly.

This poses an issue though. All the data they publish in their app must be relatable and controversial enough to matter to the general public. After a brief look at their app it becomes clear why this is problem. All their data sets end up being about gun violence, or natural disasters, or about disenfranchised people or places. Almost all of it is sad depressing stuff, or at least very polarizing. This data is what people want to see, it gives them ammo for their Facebook conversations, but it’s not so great to for graffiti to show off. Based on our conversation we figured this is where we come in. We plan to use a data set about H1B visas, which will allow us to create a much richer visualization that perhaps will be a little less gloomy.

Data collection was an interesting process. Grafiti wants to use all the data possible for as many things as possible. As a result, they were unable to give us a data set they have yet to implement. Instead, this process fell on us. The largest issue we had was the quality and quantity of data we wanted. To create a convincing visualization, we need a lot of data to distribute, but we need more than just a lot of data points. Grafiti has a handle of bar graphs and scatter plots, we need to create a more visually interesting graphic. In order to do this, we need data that has more than a few points of reference. We also wanted to use a data set that would be happy. We also quickly realized that Grafiti is made of individual people, and in order to best move forward we would have to talk to them individually rather than to 'the company' to get the best results. Initially we all agreed on a Halloween based candy data set, but upon further inspection there was just not enough there for us all to work on. We finally found a data set about immigration that we all agreed would work well. It has a lot of data on these people, from their employer to their location, and there are many interesting visualizations we can produce as a result. This is the H1B data set (Naribole).

Cleaning the data will prove to be an issue in the future primarily due to the size of the dataset. It is much larger than we need, and decided to pare down the data to what we want to visualize will be important. The main way we will clean the data is by location. Having unique locations for every applicant as well as everyone who got denied a placement at all makes this data set much harder to work with. In addition, the dataset lists job title, but with titles like 'Computer Systems' they are extremely vague and unhelpful, so that will have to removed. There is also salary data in this set, and depending on where in the country the data was entered it might be a full number or a decimal. This will have to be fixed as well. Further data cleaning will happen once an exact plan for visualization is made, which will be detailed in the final paper.

This data set is a little hard to parse due to the way it is organized. For all of the data it contains it lacks specific dates for when people were processed. As a result, they are sorted by job title. As I mentioned earlier, this metric is very broad, and that makes this data hard to read. For example, there are two jobs right next to each other, claiming the same job title, but one has a wage of 90 thousand and the next is 212.8 thousand. Not only are these jobs clearly not the same, one is an integer while the other is a decimal. Since the data has not yet been cleaned that information makes understanding the data difficult at a glance. Bigger issues lie with the employers. Due to the vague job titles, it can be difficult to see what industries are represented in the data set. It would be impossible to determine what kinds of jobs these people are obtaining.

Overall there are biggest issues with this data is not what is included but what is missing. This is an overarching problem we found with most data sets. They might have the right data but have been cleaned by another party, removing data that we would find useful. In this case it appears some important data was distilled down to a few options for easy analysis later, and that is unfortunate (Naribole).

Historically data visualization has been a means to quickly educate people. Visualizations were large undertakings and may have only been carried out when a large number of people needed to be educated of something (Halpern, 147). Since data visualizations have been used for education since the 1800's they were popularized in Europe when textbooks started to be produced for schools. They were quickly adopted worldwide and used heavily when for new inventions and discoveries. (Friendly, 21-22).

What is interesting is that by making visualizations for colleges and researchers, it still never made it to everyday life. Visualizations were reserved for papers and long readings that would be to dense to absorb any other way. Visualizations were only used because people "struggled to define [the data] with logical representation," (Halpern, 149). It was not until wealth started to be distributed after WWII ended that data visualizations were really adopted and made available to people for other uses (Friendly, 23). This is where immigration data started to pop up in newspapers and other publications. This allowed people to see trends on where and when different groups were settling down. (Friendly, 19). All of this was, of course, done by hand and data sets were kept small as a result. Large data sets that are visualized in complex ways are a much more modern invention (McGhee).

In order to help Grafiti we need to create a visualization that is unique. Since we have geographical data the best way to tackle this might be by using a map. "Data visualization, born from the marriage of classical charts and powerful computer graphics, is a way to make sense of it all," (McGhee). We need to create more than a simple map, something that illustrates not only quality of jobs acquired but also the quality of jobs rejected and the quality, in terms of salary, of these jobs in the various regions that we will determine once the data is cleaned. This means representing the data as well as having a way to describe it granularly. This visualization will have to use words to tell the qualitative story which color and shape will tell a quantitively story.

Some of the notes we received from Grafiti were to find a dataset that “Would be a great choice in terms of audience engagement and virality potential.” When initially researching we did not consider the need to focus on data that is timely. After all people still have to relate to it to care. They also encouraged us to pick a large dataset as it would help us come up with more interesting graphics and still be able to support them. They suggested a dataset that allowed for comparisons to be made inside of the data too, and warned us about having to clean a large data set. We tried to find a data set that Grafiti would love to have when presenting their idea to people, since we cannot interface directly with their app.

**Works Cited:**

Naribole, Sharan. “H-1B Visa Petitions 2011-2016.” *H-1B Visa Petitions 2011-2016 | Kaggle*, OFLC, 28 Feb. 2017, www.kaggle.com/nsharan/h-1b-visa.

Friendly, Michael. “A Brief History of Data Visualization.” *Springer Handbooks Comp.Statistics Handbook of Data Visualization*, pp. 15–56., doi:10.1007/978-3-540-33037-0_2.

Halpern, Orit. *Beautiful data: a history of vision and reason since 1945*. Duke University Press, 2015. https://slowrotation.memoryoftheworld.org/Orit%20Halpern/Beautiful%20Data_%20A%20History%20of%20Vision%20(7467)/Beautiful%20Data_%20A%20History%20of%20Vi%20-%20Orit%20Halpern.pdf

McGhee, Geoff. “Taking Data Visualization From Eye Candy to Efficiency.” *National Geographic*, National Geographic Society, 22 Sept. 2015, news.nationalgeographic.com/2015/09/150922-data-points-visualization-eye-candy-efficiency/.