CMSC 23900
Zach Kamran, Min Su Kim, Varshant Dhar

Project 1 – Write Up

Context and Purpose

The visualization serves to analyze the flows from primary subject areas emphasized in schools to the level of poverty in the school community, outcome of a campaign to fund students to the level of poverty in the school community, and further, outcome of a campaign to fund students to the metropolitan area in which the school was located[1]. The dataset for this visualization was given by DonorsChoose.org, an online charity that makes it easy to help students in need through school donations.

The visualization thus serves to answer questions such as what is the relationship between subject areas emphasized and the level of poverty in the school community? Are basic subject areas such as Literacy & Language taught more proportionally in the highest poverty areas compared to low or moderate poverty areas? Then mapping successful vs failed funding campaigns with poverty levels, we can answer questions such as: did a large portion of successful campaigns came from the highest poverty areas? As low and moderate poverty areas seem to intuitively be easier to campaign for (as they don't require nearly as much funding as the higher poverty school areas), are nearly all of them successful campaigns?

In addition, we also showed the flows between funding campaign results and different developed environments. These answer questions like: Can we compare the proportion of successful to failed campaigns between rural, suburban and urban environments?

Pre-processing

On downloading the dataset, we had to run multiple scripts in Python to extract the data that perfectly catered to our visualization – building them as csv files. This data processing was done in three scripts:

1. *CleanData.py:* Merges the outcomes.csv file (containing information on campaign outcomes) with the projects.csv file (containing information on developed environments, level of poverty and subject areas) by mapping 'projectid'. Then we removed all the data fields that are not applicable to our visualization and extracted this data into a csv file (cleanData.csv)
2. *getNumSuccesses.py*: Using the 'fully_funded' and 'primary_focus_area' fields in cleanData.csv this script counts the total number of successful and failed funding campaigns per subject area and extracts this data into a csv file.

[1] https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data

3. *getAllStats.py:* Using the 'poverty_level' and 'school_metro' data fields along with the data fields in *getNumSucccesses.py*, we found the number of successful & failed campaigns in each poverty level and further the number of successful & failed campaigns in each metropolitan developed environment.

We then divided up the data manually into csv files that perfectly catered to our visualizations. These csv files are failed_by_metro.csv, poverty_by_success.csv and focus_by_poverty.csv

How does it work?

We decided to use three two-channel Sankey diagrams to explain the flows between different correlated data fields in the dataset. We initially built a two-channel Sankey diagram with subject areas and poverty levels to answer questions about subject area emphasis and how the proportion of emphasis differed between schools in different levels of poverty. We wondered if students in the highest poverty areas were being exposed to subject areas like Music, History, and Sports and whether schools in moderate to low poverty areas had any emphasis on basic skills like Literacy. As a result, we thought the best way to highlight the flows of subject areas between different poverty levels would be through a Sankey.

On enjoying the outcome of our initial visualization, we decided to visualize another Sankey representing the flows from the funding campaign outcomes to the poverty levels, as it would then tell us how the projects in different levels of poverty fared. In showing the flows between different levels of poverty and campaign outcomes, we were able to determine what proportion of the campaigns from different levels of poverty were successful and failures.

Based on this new channel of campaign outcomes, we also decided to add a third Sankey that showed the flows from the funding campaign outcomes to the different kinds of developed environments (urban, suburban and rural). We were thus able to answer questions related to the correlation between successful and failed campaigns and the different development areas in which the schools were located.

Although the campaign-poverty Sankey and campaign-metropolitan area Sankey represent different flows, we decided to synchronize the color schemes we employed for the shared data variable of funding campaigns. Using shades of blue for successful campaigns and shades of green for failed campaigns, we were able to connect the two flows and notice trends that existed with campaign outcomes in different metropolitan areas and different poverty levels. The different shades distinguished the different metropolitan areas and poverty levels, while the different colors distinguished between successful and failed campaign outcomes.

For the subject-poverty Sankey we decided to employ a rainbow color scheme that would distinguish each of the different subject areas while showing the flows to different levels of poverty. Unlike the other two diagrams, we decided not to use shades of color while mapping the data to different poverty levels. We thought this would create extra confusion while

employing a channel with seven data values (Literacy & Language, Math & Science, Music & The Arts, Applied Learning, History & Civics, Health & Sports, Special Needs) instead of our other diagrams that had only two data values for the starting channel. The rainbow scheme was also randomized to show that no order existed between the different subject areas – the only emphasis was on differentiating them and their respective flows.

In terms of the layout of the visualization, we took Andrew's advice and decided to put two Sankey diagrams that were connected to each other right on top of each other, and decided to rotate our separate subject-poverty diagram, stretching it out and placing it on the side. In addition, we thought the shadow boxes on the nodes made the Sankey more legible and the flows clearer.

Alternatives and arbitrariness

Our initial idea was to build a visualization with a color scale representing the poverty level, with each school represented by either a triangle or a circle on a map visualization of the United States. This was because the dataset also consisted of latitude and longitude coordinates. We decided against this idea as map visualizations are very cumbersome, and mapping poverty levels within different areas in a large map of the country wouldn't make visual sense. Poverty levels vary immensely within cities, and so, a map of the country would be very confusing.

When we then decided to do a Sankey diagram, we first thought of building it by doing a two-channel Sankey with the state in which the school was located as one and the outcome of the funding campaign as the other. However, the number of states compared to the number of outcomes (two) made the diagram look disproportional. And so, we chose to replace the states channel with subject area (seven) – making the layout look far better.

We also thought of creating one large Sankey diagram that covered all 4 data variables in the visualization, but we had trouble with the Sankey API and were under a time constraint when this idea occurred to us. Furthermore, we would've had to mine the data again looking for counts that covered different possible poverty levels, metropolitan areas, funding outcomes and subject areas. As a result, we chose to create three separate Sankey flow diagrams that addressed the questions we wondered the most about this dataset.

Who did what

Varshant Dhar: Found the dataset, wrote the write up.

Min Su Kim and Zach Kamran wrote the code for the visualization and researched possible visualizations. Min Su in did research into colorings for the sankeys. Zach Kamran did research into the visualization methods. The code was written collaboratively by Zach and Min.