

Project 2: Kickstarter Dashboard

Jaret Noffsinger, Logan Mancuso, Blake Pathammavong, Kevin Zheng

April 29th, 2019

Code: <https://github.com/JarNoff/CSCE567-Project-2>

**Introduction**

Our project 2 was based on a dataset of the website Kickstarter. This dataset contains all campaigns, up to March 14th, 2019. There are roughly 418,000 different campaigns of all categories contained in our 1.7GB CSV data file. The dashboard that we have created explains where people that create campaignsto be funded are from, the geographic distribution of various categories, the best place to start a campaign in a given category, and seeing what percentage of these campaigns meet their goal. This dataset contains all campaigns, up to March 14th, 2019.

**Page 1**

The first graph shows the geographic distribution of categories around the world. The visualization shows the most recent 1,500 campaigns because with Power BI, it only allows 1,500 without a valid ArcGIS Power BI tools license. If you hover over the location dots, there is a tooltip will show the variation of campaign categories.

**Page 2**

The second graph shows the amount pledged by category and location. There is a list of locations in the bottom right to select. The data will show the category with the most funded to the least funded. If you hover over the bar you can see the average amount pledged to that category in that selected location. The gauge graph in the tool tip also shows if the average campaign in that category met their goal, the average amount pledged, and the max amount pledged.

**Page 3**

The third graph shows that categories in a word cloud; the larger the size of the word, the more campaigns in that category. The slider on the bottom right allows the user to select the minimum and maximum amount pledged. If you increase the minimum amount pledged to somewhere around $20,000 USD, the category games seems to have the largest number of entries all the way up to the maximum value, $8,596,474.58.

**Page 4**

The fourth and final graph shows the best countries for each category. The circular graph to the right displays the standard deviation of pledge amounts for each country in the currently selected category. The country with the highest standard deviation of amount pledged in the particular category chosen in the best country to create a campaign for the selected category.

**Data Wrangling Techniques**

Originally, we were using the CSV data file, at an approximate size of 1.02GB. The problem with this was that there were objects nested inside of certain fields, which was messing up the delimiter (objects inside fields contained commas too). To remedy this situation, we found a CSV Parser library on GitHub. We then wrote a program designed to parse through the csv data file, and “un-nest” the nested fields.



Figure 1: An example of nested objects in the CSV file

So for example, in the above picture, in line one, the program would extract the value “music/rock.” All of the code for the CSV Parsing program can be found on GitHub; the link will be on the cover page.

Once the program was almost complete, we saw that the source we got our data from had a JSON format. So instead of finishing fixing the bugs in the CSV Parser, we downloaded the JSON file and flattened it via the command line in UNIX. This resulted in a data file that was correctly formatted (for the most part), with every possible field. This file was about 1.7GB, so a little larger, but made the data much easier to read.

After parsing the CSV file, we discovered that the categories were copied over in duplicates/triplicates for a few cases. So, for example, the categories “music/rock” and “Rock” where copied into the data file. To fix this, we used the Power BI group functionality to group categories that were duplicated or triplicated together, so we got all relevant campaigns when choosing by category.

Another issue was that the Kickstarter data is that it contains two letter country codes, instead of three letter country codes. This was a problem because Power BI does not formally support 2 letter country codes. To fix this, we imported a data table from Wikipedia entry of the ISO 3166-1 standard. This standard establishes codes for the names of continents, countries, and special geographic areas. Once we imported this data, we used the Power BI model editor to mark the location column in our dataset as a one to many relationship with the country code column of the new table. This solved our problem and gave us displayable location names that worked with Power BI.

**Why Power BI?**

We used Power Bi because most of us were not adept in CSS coding and styling; when we saw that Power BI was based on a visual, drag and drop style, similar to Tableau, this made us feel much more confident when formatting and displaying our data. Power Bi can read data from JSON files, CSV files, directly from the web, and more, giving it the advantage of extensive database connectivity capabilities. Power BI utilizes the same Microsoft user interface environment as Windows 8 and Windows 10, which helped us become familiar with the software much more quickly. Power BI is also has strong tools for predictive modeling, reporting, and optimization, and, in addition, is free for all U of SC students.

However, Power BI is definitely not the perfect platform for all applications. We ran into many small problems while creating the visualization. For example, we found that we could not modify the text displayed in the tooltip; the only way to manipulate this text was to change the data directly. Another example of this is trying to group the locations by countries or continents. Power BI does not support sub lists, so we could not organize or sort the list of locations by continents in the visualizations. Power BI lacks a lot of the custom functionality we could have, if we wrote the dashboard from scratch with Python.

Overall, Power BI is a very intuitive, user friendly application for creating data visualizations, but it has a good deal of limitations and drawbacks. Power BI is not likely to be a “one stop” program for manipulating data, but if the data is already formatted, it is one of the quickest tools for creating simple graphs.

**Works Cited**

Web Robots. (2019). Kickstarter Datasets (2019-03-14). Retrieved from https://webrobots.io/kickstarter-datasets/