DIVE INTO DATA: WHAT'S COLLECTING IN OUR WAVES & WATERWAYS

PROVIDING DATA SCIENCE RESOURCES TO NON-PROFIT ORGANIZATIONS

Michael Campellone, Ikkei Itoku, Mia Zhao, Anna Cianciara, Nicole Barberis, Eduardo Hermesmeyer

Bloomberg

ABSTRACT

As non-profit organizations continue to face challenges in backing their respective initiatives to drive policy change and receive the necessary investment from both government and private sector sources, the increasing reliance on data to support their claims becomes fundamentally critical. That being said, non-profit organizations often lack the necessary resources internally to perform the required data science approaches and techniques to fully leverage the power behind the data the organization is collecting. With Bloomberg's existing commitment to supporting philanthropic initiatives across the Public Health, Environment, Education, Arts and Government Innovation spaces, we as an organization want to ensure non-profits have access to the resources crucial to applying data science techniques to data that will ultimately drive policy and behavior change in the future.

INTRODUCTION

In order to pilot this skills-based volunteer initiative, we partnered with Clean Ocean Action (COA), an organization Bloomberg already had an extensive relationship with. Bloomberg has had a long time history of performing Beach Sweeps with COA since 2011. As of July 2015, 185 unique employees have dedicated almost 800 hours across 17 volunteer events.

SCHEMA ARCHITECTURE

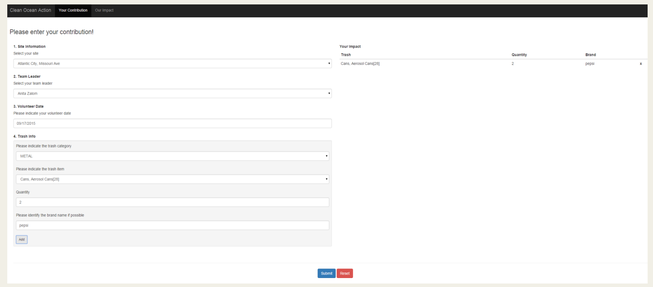
The first task of introducing the new data governance system was to thoroughly review the current data retrieval process in order to fully understand the data flow from beginning to end, as data could be deteriorated at any point of the process. To begin, we interviewed the COA and American Littoral Society staff as well as internal Bloomberg employees who participated in the COA beach clean-ups to identify the steps volunteers take to report the collected trash items from beaches and the procedures the COA staff follow to compile the data. Based on the observations and analysis of the 1993-2014 datasets, database schemas were developed in MySQL, for both its reputation and significant presence in the data science community.

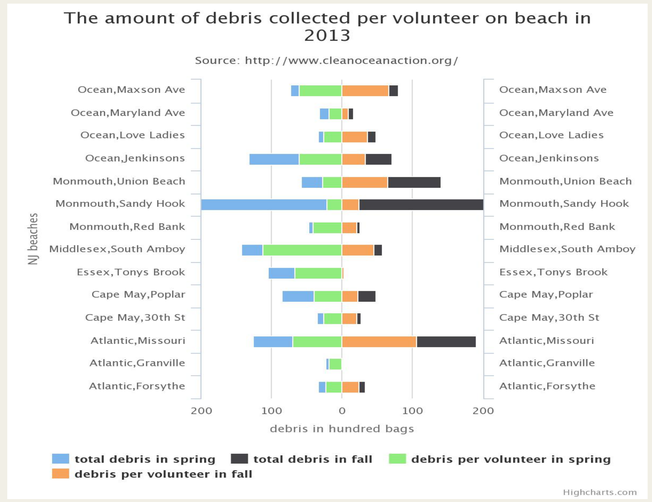
HISTORICAL DATA CLEAN-UP

After designing the schemas, data munging was required. Due to the data volume and inconsistency, this process demanded a significant amount of time. Each year contained two Excel files with multiple sheets of detailed data collected by COA. Furthermore, there was little format consistency between files and the classified categories had changed overtime. Thus, there was no way to programmatically process these files and manual examination of each category and file was needed. Utilizing volunteer time at Bloomberg, the most recent years of the data were standardized and successfully transferred to the database. The remaining files will undergo a similar process during an upcoming "Bloomberg Datathon" event.

WEB APPLICATION

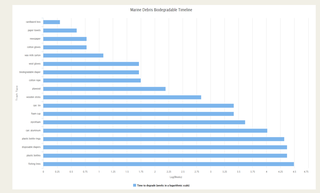
After the schema design and significant data munging, a new data entry framework was needed for volunteers and COA to update collected items into the database. To solve this, we created a custom web application based solely on open source solutions so that COA could invest the resources for its core operations. For readability and future maintenance purposes, Python was our language choice and we utilized a micro web framework, Flask, as the backend. The front end was built on top of a popular open source solution, Bootstrap, so the site could be optimized for desktop computers and smartphones. This application will not only solve data entry and integrity problems but will function as a real time data analytics dashboard as well. As of September 2015, the application was successfully deployed from an internal Bloomberg web server to Amazon Web Services, more precisely, Amazon Relational Database Service and Elastic Beanstalk.

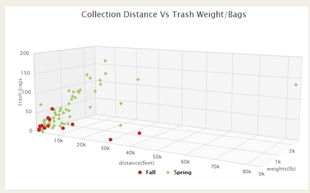




DATA ANALYSIS METHODS

Using the semi-annual (Fall and Spring) trash data provided by COA, we analyzed the volume of trash collected against: macro indicators including GDP, unemployment rate, and New Jersey population; related industry indicators including aggregated inventory, sales, and revenue data; and finally, related company information such as Pepsi and Coca-Cola revenue and sales data from the US market.
We performed Granger causality test on the data series containing strong correlation, setting our constraints to $t<0.05$ as statistically significant and r^2 to 0.8.





CONCLUSIONS

The development and implementation of the web application, as mentioned above, will serve four purposes. The first will allow our Bloomberg team to standardize the remaining historical years of data. Second, the application will allow volunteers and COA to standardize the data they are collecting and directly load this data into the database. As a result, future data will no longer need to undergo the time intensive process that was necessary for the historical data we received. Third, we see the application serving as a tool to increase volunteer engagement. Analytics on the web application are updated real time and allow an individual or team leader to visualize the impact they are making. Lastly, the analytics made available to the organization via the web application will allow COA to plan more efficiently from an operations standpoint. We see this application being used to properly allocate resources at the various beach sites and also to allow the organization to reflect on prior Beach Sweeps and strategies that may or may not have worked.