Final Report

SI 485
April 19, 2021

Puente

Hamza Baccouche
Katherine Berry
Drue Froeschke
Mattias Siimar

Overview of Project

We are working with Puente Desarrollo Internacional (Puente), a global non-profit organization that uses data collection and analysis technology to tackle some of the biggest challenges in developing countries. Puente's mission is to connect international development organizations to local institutions that represent communities in need, to make development more efficient, collaborative, and sustainable.

The overall goal of the project was to provide Puente with extensive data analysis. This analysis is largely based on clusterings of the various communities they work with in the Dominican Republic. The goal of these analyses would be to provide Puente with information that would show them which communities are "high need." We aimed to provide Puente with analyses that would become one of the many ways Puente identifies communities to help, and it is now one contributing factor.

Project Deliverables

Our final product for this project is visualizations and analysis. To do this analysis, we used unsupervised machine learning to cluster the data based on province, then at a more granular level by community. We then used these clusters to provide Puente with in depth analysis on many health-related factors affecting residents in the communities they work with. These factors include, water access and type of water drank, trash disposal access and frequency, most prevalent injury types, biggest overall problem affecting the community, etc. We then summarized takeaways from these graphs, for instance, lack of filtered water might be related to high rates of stomach/intestinal related injuries. After the analysis was done and the graphs had been made, we changed all the graph color schemes to match Puente's brand guidelines.

As an additional component of our project, we built Puente a data cleaning function. Data cleaning was a large chunk of the work we put into this project as the data was very messy with missing values, misspelled words, etc. After we finished cleaning the data so we could begin to use it, Puente asked us to build a function that would clean their datasets in the future. The Puente team has also spent much of their time working on data cleaning and thought a function that automated this process would be very helpful. Therefore, we took what we did in the data cleaning process and created a function that would do everything we did.

All of our code, visualizations, and documentation can be found in the project Github repository (https://github.com/hamzabacc/puente). The folders .ipynb_checkpoints, Cleaned CSV Files, Clustering Starter Files, and JSON Files include the files we received from Puente and some of the messier exploration and data cleaning work we did. We decided to include these so that Puente has everything and their data scientist could look through it at our leisure if they wanted to explore our processes more in depth. The EDA.ipynb file includes some rough Exploratory Data Analysis that we used to help find our direction and get a sense of the data. The datacleaning.ipynb file includes our data cleaning processes and the data cleaning functions we created. The analysis_code.ipynb file includes all of our analysis, and automated visualization function that creates visualizations for any variable Puente would like in the style of their brand (including font and colors)

Approach

Our project has two primary audiences: Dominican residents and the Puente team. For the former, we set out to convey the most important data to Dominican residents in a way that is understandable at a 4th grade comprehension level. For the latter, we evaluated numerous approaches to machine learning in order to identify communities of need, and settled on an unsupervised clustering model.

The first step in making the visualizations was to determine which data points were most important. In keeping with the bigger picture of the project, we identified a handful of variables that were strongly correlated to the wellbeing of communities, such as access to clean water, whether trash is picked up or burned, and frequency of medical checkups.

The clustering model was a more extensive task. We worked through identifying the optimal number of clusters such that we can balance the boundaries between communities with the need for inter-region division. This can help identify areas of high need within communities that may otherwise look to be well off and are thus overlooked.

Our process involved many intentional pauses to step back and look at the bigger picture to ensure that our variable selection, model selection, and cluster volume selection all moved us closer to our goal of identifying high-need communities that Puente can help.

Value Created

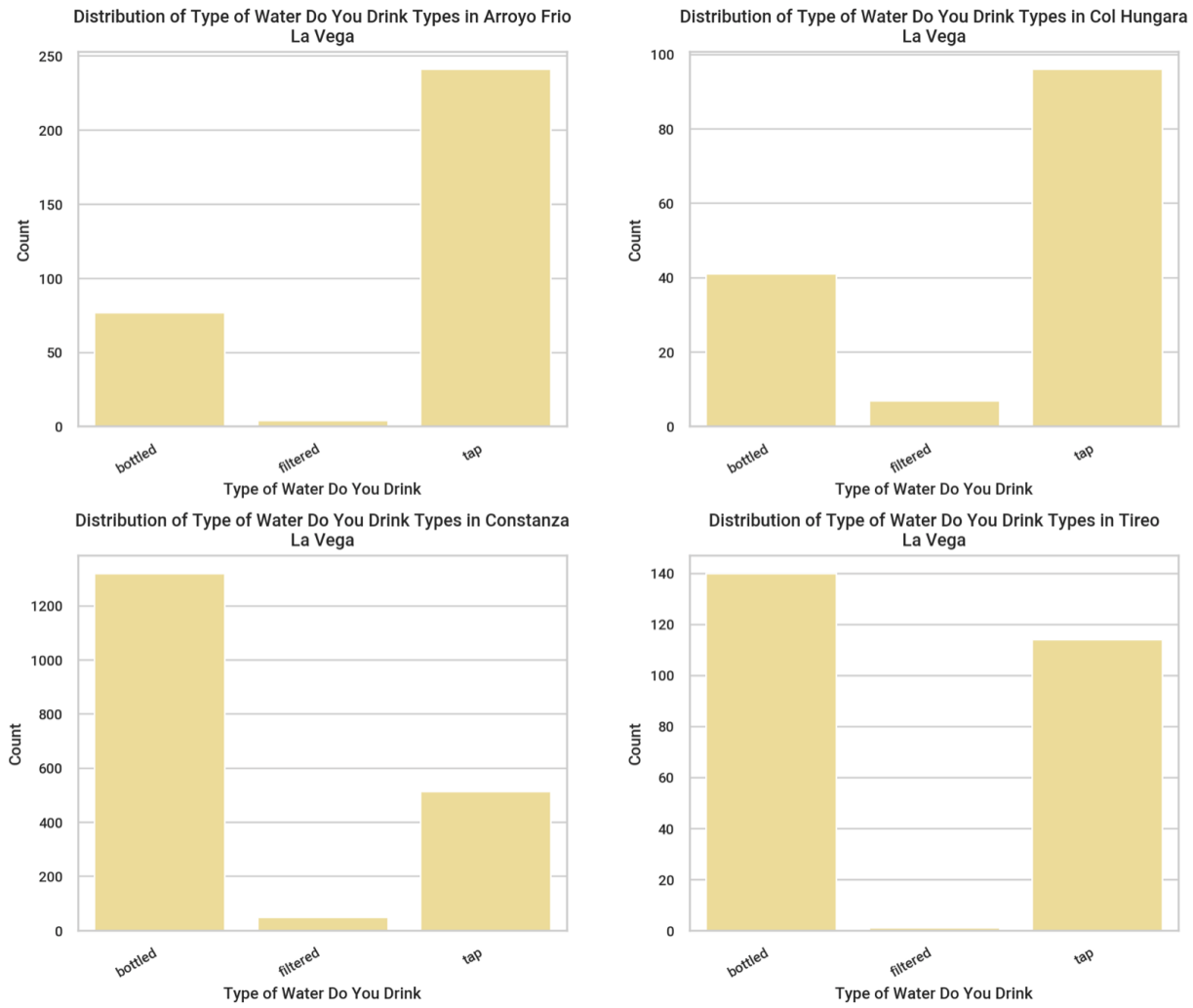
We provided Puente with our Github repository that has extensive data analysis, including over 60 visualizations to easily depict our findings throughout the past year. An important part of this project for us was usability, so we created a data

We have provided graphs on topics of water access, type of water residents have access to, trash disposal access, injury type, bathroom/latrine access, the biggest problem a community is facing, etc. We also made these graphs aligned with Puente's brandbook so these graphs are ready to use for Puente in any client facing situation as well.

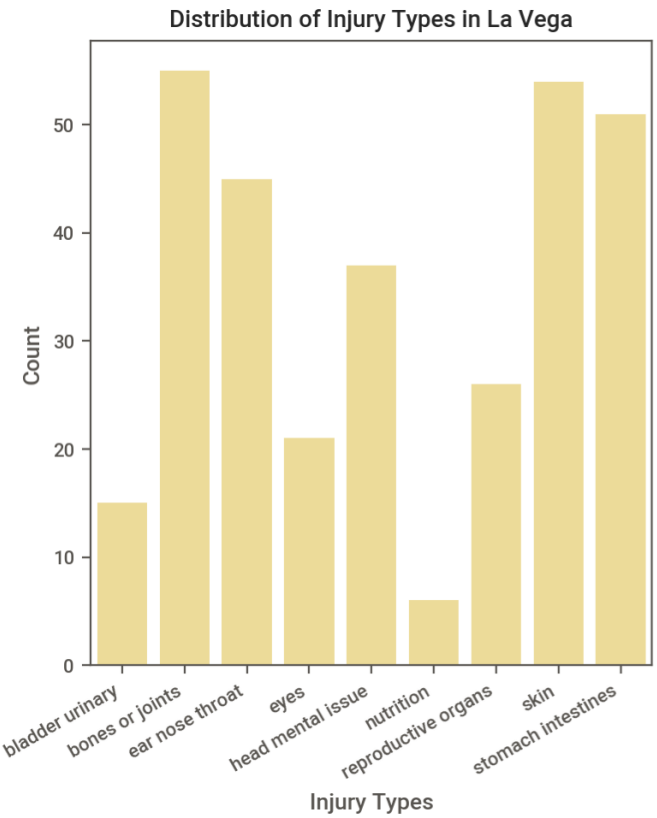
Throughout the Github we made sure to include comments in our code and explanations on our jupyter notebooks so that both technical and non-technical members of the Puente team can easily understand all steps we took to produce our outputs and graphs. We also provided Puente with a data cleaning function they can use to automate data cleaning in the future, which was a large portion of the work we contributed over this past year.

Lastly, we found two main takeaways from our visualizations that we thought would be helpful for Puente, which are highlighted below.

Finding #1

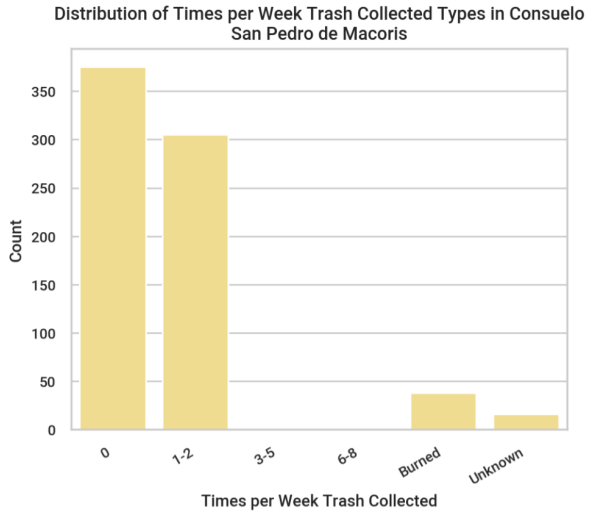


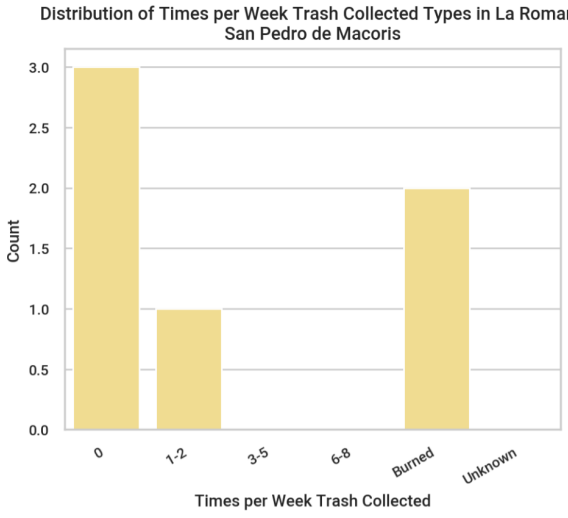
We can see that a considerable amount of people in La Vega province drink tap water which tends to be dirty in many developing countries. Drinking unfiltered water can lead to hazardous health issues such as diarrhea, cholera, parasites, and other intestinal problems.

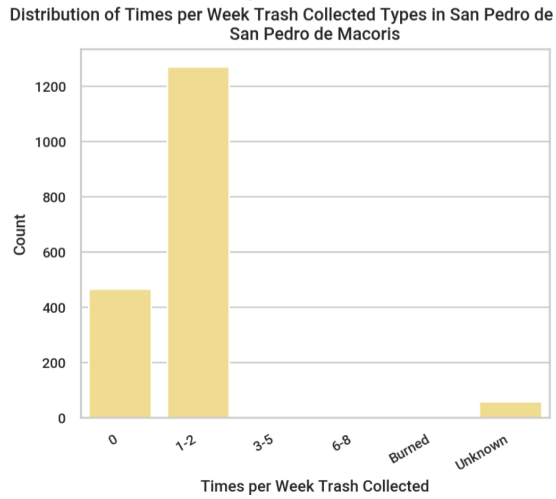


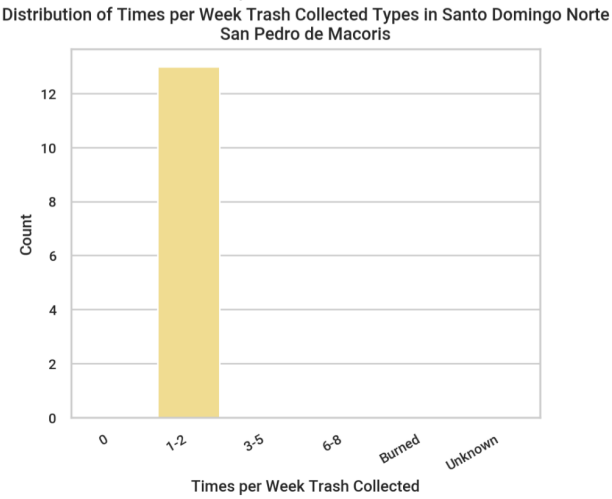
The visualizations above allow us to see that people in La Vega province have intestinal issues which could be caused by unclean water. Based on the finding, we can suggest that these two factors could be correlated, and Puente should conduct some water tests in communities in the La Vega region to see if more water filtration systems would help these communities avoid several health issues.

Finding #2

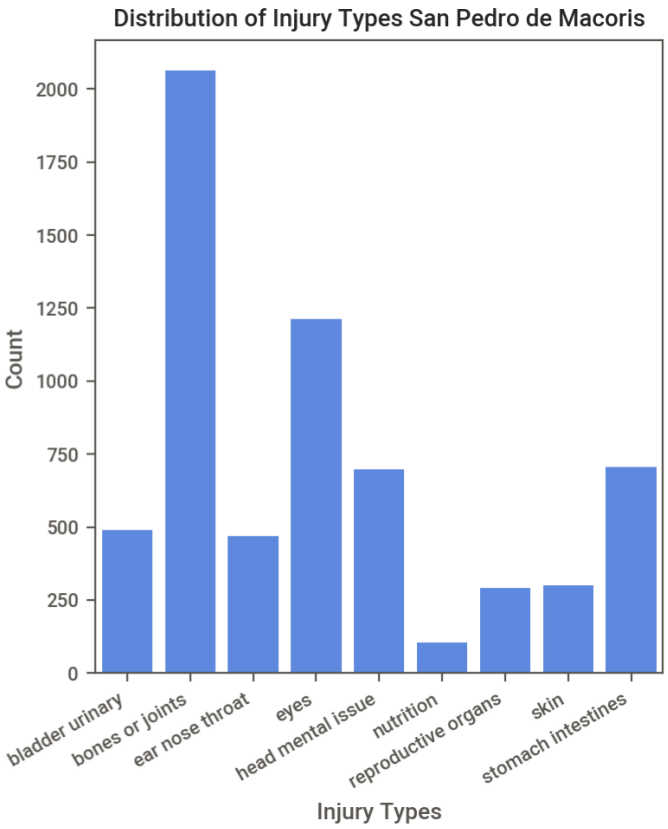








From the graphs above, we can see that many residents in the San Pedro de Macoris community do not have frequent access to trash disposal, with many residents saying their trash is never collected, with other residents even reporting their trash being burned due to lack of collection. Leaving trash sitting out for long periods of time and burning it can lead to health conditions related to eyes, and ear/nose/throat.



The visualizations above allow us to see that people in San Pedro de Macoris province have eye and ear/nose/throat issues which can be caused by lack of trash collection. Based on the finding, we can suggest that Puente should focus their attention on the several communities in the San Pedro de Macoris region that do not have frequent trash collection, to see if they should implement more trash collection in order for people to avoid several health issues.

Relevant Correlations

In an effort to find more concrete takeaways, we ran several different correlations to indicate potential areas of high need. For reference, an r value indicates the correlation coefficient, and it ranges from -1 to 1. When r is close to -1 or 1, the more closely those two variables are related. An r value close to 0 indicates that there is no significant relationship between the two variables.

- The correlation between condition of cement floors and condition of the roof in a house, meaning if one is poor, the other typically is too, is $r \approx 0.4$.

- The correlation between years lived in a given house and years lived in the community, meaning people usually stay in one house and don't move houses often, is $r \approx 0.65$.
- There is a much stronger correlation between where you go for medical problems and where you go for dental problems for those who answered Ramon Santana Hospital ($r \approx 0.85$) vs, other responses ($r \approx 0.5$), meaning those who go to Ramon Santana are much more likely to go there for both medical and dental services, whereas other respondents aren't as likely to go to the same place for medical and dental services.
- The correlation between having no water access and believing the biggest problem of your community is water, meaning people who don't have water access believe that to be a bigger problem than others like crime, infrastructure, lack of healthcare, etc. when those other problems also occur, is $r \approx 0.45$.