Prison Data Collection, Cleaning, and Visualization with D3, R, and C#

Alexis Saltzman
Catherine Krueger
Christian Cleberg
Joel Wach
Meghana Kurupalli

Group 2
CSCE 320/411/811

1. Introduction

A large proportion of the world's prison population is in only a few countries, primarily the United States, Russia, and China. By analyzing data of the locations of prisoners in all countries, this project intends to make the information easily understandable and accessible, while looking for trends in the data that can be seen in a heat map. What do the countries with low prisoner populations have in common? Does it appear that areas surrounding the highest have similar levels? What factors might contribute to areas having high incarceration rates? This project proposes to create data visualizations that will make this questions easy to understand and learn. This project aims to aid in the learning and understanding about possible answers to these questions by visualising the data we have.

2. Challenges

We expect to encounter a few challenges in this project.
Firstly, collecting the data is a real challenge. There is no database readily available to work with. We had to go through some manual and laborious processes to obtain the data. Then, we needed to ensure the visualization can be easily updated when new information is given. Also, we needed to choose a language and project structure to build a data analysis application that is fast and space-efficient. And finally, we needed to create a visualization that is detailed enough to make a lot of information available without sacrificing the user's ability to understand the data.

3. Related Work

There are studies on the prison population in different parts of the world[3] and statistics of prison population of all the countries in the world[2][1]. Many studies have focused on the United States for having a very high number of prisoners per capita. The studies in the United States delve very deeply into the underlying causes of these high incarceration rates - including the law, the police force, poverty rates, and possible racially charged motivations. We aim to understand on a global scale what causes higher incarceration rates - what certain countries may be doing to have fewer incarcerated citizens, and we believe we can accomplish this by using a data visualization tool.

4. Approaches

We approached the problem to gain insights on the data, clean the data, and combine multiple data sources to create a heat map of the world to show trends in incarceration as related to other factors. We want to consider the following factors, in every country, for this project :

- Prison population
- Rate per 100,000 citizens
- Female rate
- Juvenile rate
- Occupancy level
- Type of government.
- Number of institutions

We used D3.js to create this visualization, and C# for the backend to process and format the data for D3. Two data sources (one CSV and one JSON file) were used that each provided slightly different fields and information which was combined to give the final result with the fields above.
As an additional source of insight, we performed some data analysis with R as well. We imported the JSON data file and coerced it into a dataframe. The World Prison Brief dataset was loaded into the workspace and using these two datasets, both separately and merged together, we were able to experiment with analysis and visualization. This allowed us to familiarize ourselves more completely the data and fine-tuned our questions. We chose to graphically represent the answers to our questions using the ggplot2 library[5], and explore the variable importance defined by the K Nearest Neighbor algorithm using the caret package to train the algorithm with the data.

5. Expected Results

We expected to build an application that would create data visualizations to identify trends in worldwide incarceration. We expected these visualizations to make clear certain correlations between incarceration, government styles, and other factors. Finally, we expected to generate some insights from the use of R. For example, were were anticipating that the KNN results would prove that

government type had a high correlation with prison rate size.

6. Results

Our project collected data from multiple sources. With this data, we generated these insights from R:

- for most countries, and definitely the countries with the highest rate of prisoners, the rate has be growing steadily over the last ~dozen years.
- The rate of prisoners overall correlated with the rate of imprisoned women and juveniles.
- (insert about rate of institutions vs prison size
- The government type was not among the variables that KNN deemed important to training the algorithm, it appeared to be a very weak correlation.

Additionally, we used C# to clean and combine two data sources, which created a successful visualization in D3.js.

7. Evaluation of Results

The accuracy, precision, and recall of the data returned to the user is our first priority. We successfully cleaned and combined the data sets, and tested the webpage created to ensure that the right information was returned, and the wrong information was not. Our second priority was to be able to create visualizations to help answer questions about incarceration rates in countries around the world.

Example cases that can be answered by our D3 and R visualizations and analysis include:

- What countries have juvenile prisoner populations greater than 5,000?
- Do theocratic governments have higher rates of prisoners than monarchies?
- Is the rate of female prisoners in a population similar or dissimilar to the rate of male prisoners?
- Is the rate of juvenile prisoners in a population similar or dissimilar to the rate of adult prisoners?
- Does it appear by the data visualization that k-nearest neighbor would be an adequate predictor in prison population rates?

We successfully met all of our goals as we collected data from multiple sources, gained insights from R, and were able to clean and combine the data into a CSV for use by D3. From this project, we learned about data collection and what makes data valuable and compatible with other data - knowing what columns are necessary and what columns can be joined on. We also learned how to combine and clean data with C#, which entailed learning to read both json and csv files as models, cleaning both datasets, and then learning to combine those models into a final format. Finally, the results from the K Nearest Neighbor were not as successful in providing insights as we had hoped. For example, we hypothesized that there would be a stronger correlation between rate and government type, and there was not. We believe this is due to two reasons: 1)The data had enough NA values mixed in that there were not enough points to allow the algorithm to train thoroughly, and 2) there were approximately 20 different types of governments, and when the dataset is not large the algorithm doesn't get a chance to give classification significance to each. The graphs and plots made with ggplot2 gave us a deeper understanding of the way the data points relate to each other.

References

[1] "World Prison Brief." Norway | World Prison Brief, 2016, www.prisonstudies.org/.

[2] Walmsley, Roy. *World prison population list*. London: Home Office, 2003.

[3] Berman, Gavin, and Aliyah Dar. "Prison population statistics." *London: House of Commons Library* (2013).

[4]Reference article at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198040/

[5] CRAN (The Comprehensive R Archive Network) Package repository https://cran.r-project.org/web/packages