Exploring 2022 Indian NGO Distribution

DaanMatch is a nonprofit that focuses on getting resources to the smallest non-governmental organizations (NGOs) in India with a goal of educating ourselves and others about how nonprofit resources are distributed in India. Currently, we are trying to address the social inequalities in fund distribution. If we can understand what the bigger NGOs are doing, then we can understand what the smaller NGOs need, what resources they need, etc.
To learn more about DaanMatch, **click here**.

Authors

Spring 2024 Student Researchers: Dhruva Bhagwat, Adam Mahmoud, Breanna Cabusao, Justin Lee, Siyi Huang, Scarlet Shi, Esther Yoo, Cole Yap, Yash Natu, Rahul Sathyanarayanan
Advisors: Kenneth Kron, Cara Foss Arellano, Patrick Guo

Affiliations

We are affiliated with the University of California at Berkeley as undergraduate students through the Data Science Discovery Program. We are also affiliated with the nonprofit organizaton DaanMatch, the hosts of this project.

Background & Significance

Recent data reveals a concerning trend in India's organizing funding allocation: the majority is directed to urban projects which primarily benefits multinational non-government organizations (NGOs), yet about 70% of NGOs receive disproportionately low funding as they are community and rural-based. Despite the initiation of corporate social responsibility (CSR) funds since 2014, which are designed to bring companies and NGOs together and support them, they tend to favor highly visible urban NGOs, neglecting grassroots organizations in rural areas. This disparity underscores the urgent need for equitable distribution to harness the full potential of CSR in addressing local needs and fostering sustainable development nationwide.

Abstract

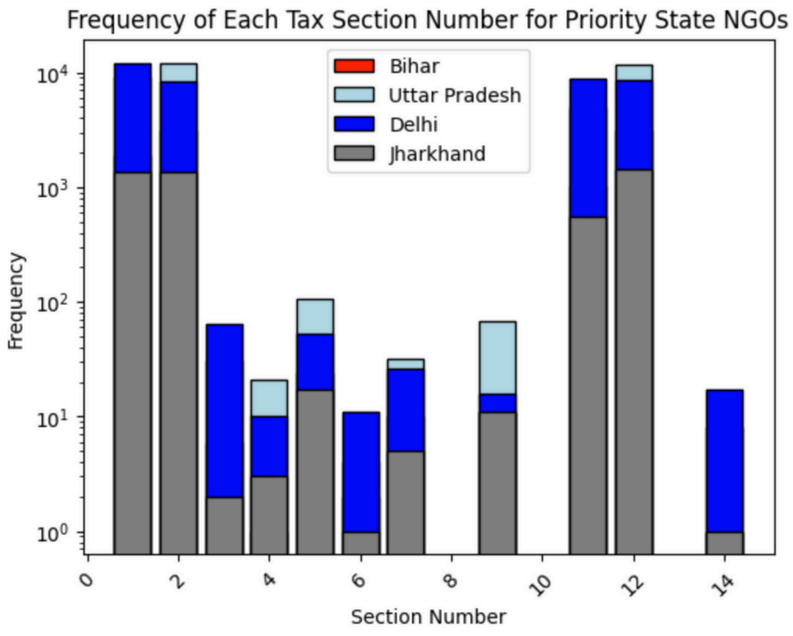
Our task involved cleaning and merging the tax records data, standardizing formats and conducting exploratory data analysis to identify relationships between institutions, analyze distribution patterns and gain insights into the tax sections used by different NGOs. The purpose of our study is to leverage tax record data to inform decision-making strategies within the nonprofit sector, particularly focusing on addressing social inequalities in fund distribution in India.

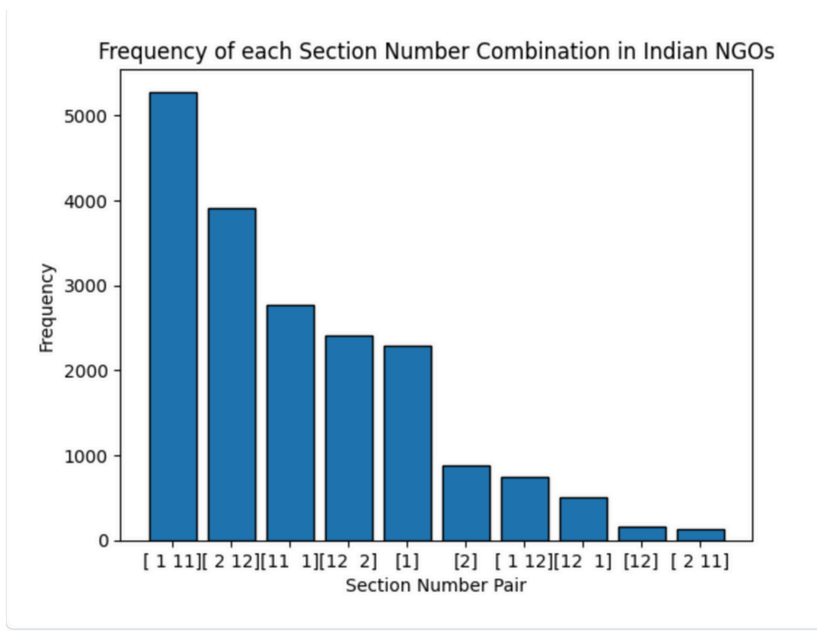
Cleaning/Merging

- Tax Records dataset is 100000+ lines, only interested in 4 priority states: Bihar, Jharkhand, Delhi, and Uttar Pradesh.
- Cleaning Process
  - Filter for 4 priority states.
  - Replace all null or invalid values with the word "NULL"
  - Convert all registration dates to date-time format for consistency
  - Title-case all address names for consistency
- Merging Process
  - Convert all section number strings to their respective integer
  - Group by Pan-ID and return the array of all Registration-IDs and Section Numbers corresponding to a specific institution
  - Join the grouped columns for Registration-ID and Section Number to a grouped table by Pan-ID.
  - Drop duplicate columns and rename columns to their original names

Tax Section Number EDA

- Exploratory Data Analysis on the different Tax Section Numbers within our Dataset
- Creating New Data Frame
  - Utilized Dataframes titled 'Bihar', 'Up', 'Delhi', 'Jharkhand' to create a new DataFrame
  - Retrieve the section number for each row in each table
- Creating Visualisations
  - Visualisation to see Frequency of of each Section Number
  - Visualization to see Frequency of each Section Number combination





Standard Address Formatting

What is the Standard Address Format (SAF)?
In order to have uniform addresses for each NGO, we conformed the inconsistent address data that was provided to us via the tax records data into the format given below:

SAF	Example NGO
Recipient	**Kashi Yoga Evam Mulya Siksha Sanstha**
House Number | Street Name	**27 Khojwa Bazar S.o**
Dependent Locality	**Varanasi**
Double Dependent Locality	**–**
Locality - Postal Code	**Chhitupur - 221010**
India	**India**

How did we achieve this formatting for all NGOs in the priority states?
We created a function that takes in a Pandas DataFrame (containing the desired NGO), and an institution's name or pan-id (unique ID assigned to each NGO) as arguments...

```
def formatted_address(df, institution="", pan_id=""):
```

...and prints that institution's address in the standard format.

Conclusion & Next Steps

By thoroughly cleaning the Tax Records data and coercing the addresses into a standard format leaves for more efficient data analysis to be done in the future. Our next step in the following semester would be to find the exact geographic location utilizing geospatial analysis to pinpoint the proximity of the organization to the most profound need. In other words, we would be using the data to find where the closest nonprofit is to help with the most productivity.