St. Joseph's University

Bengaluru, Karnataka - 560027

Hackathon Final Report

Submitted by:

Yobah Bertrand Yonkou (222BDA15)
Alvina Joanna (222BDA63)
Saba Santhosh (222BDA65)

Submitted to:

Srividya Suresh
Adjunct Professor
Department of Advanced Computing
St. Joseph's University

Aim

The world we live in is diverse to its very core. Our nation, India is a beautiful representation of the diversity that exists not just in nature but also among people and their lives; encapsulating the spirit of 'Unity in Diversity'. Given that diversity is our strength and not our vulnerability, we must seize the opportunity to appreciate and take note of the same using our publicly available national census. In our work for the Hackathon, where we were challenged to find a problem statement and provide a relevant solution - we took it upon ourselves to find aspects that could best express the most prominent characteristics of our population and then calculated their diversity in each state. This analysis helps with giving intuitive indices that will help understand the richness of diversity from quantitative data.

Domain

Demographics analysis, Cultural diversity indexing.

Problem statement

Computing diversity index for occupational classification, religious communities, language groups and migration patterns from 2011 Indian Census datasets.

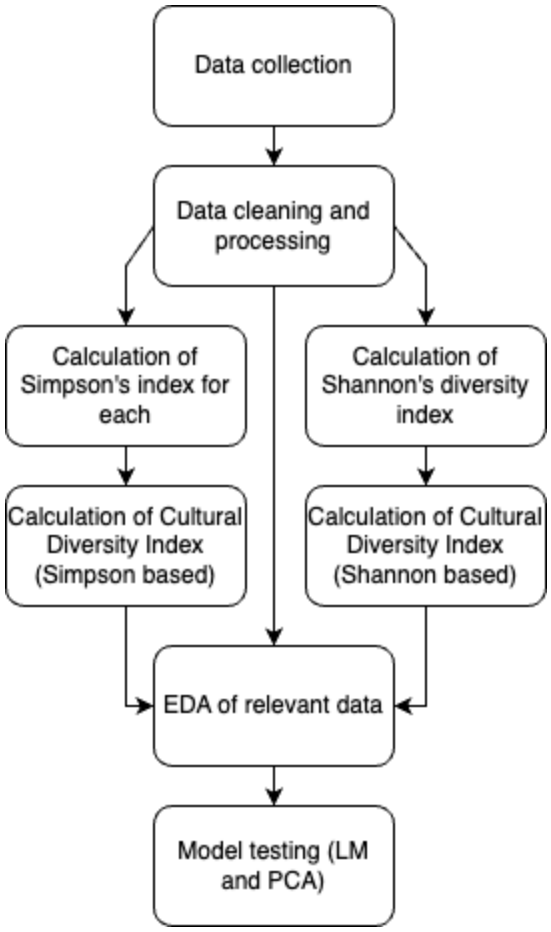
Introduction

Our first dive into the datasets available for analysis during this Hackathon was to understand what information was available to us and how that information could be analysed to gain new or interesting insights. As we sifted through the data, we found many datasets that contained multiple classification and levels of data - showing a depth of information that could be extracted to infer meaningful insights. That was when we got the idea to find the diversity of the following aspects for the different regions in India:

- Occupational diversity: shows how diverse is the population in terms of profession/occupation as this will show a greater mingling of people from various classes and sections of society
- Migration diversity: helps quantify how much of a melting pot of cultures exists within a region
- Linguistic diversity: with an inflow of people from various regions, the linguistic diversity also increases and the people of the place are more sensitive to other cultures and sentiments
- Religious diversity: an important index to show an intermingling and co-existence of people with different belief systems

Our journey into the data and its analysis enabled us to learn that the field of diversity indexing stems from the field of ecology and environmental biodiversity. We employed the formulas and were intrigued by the results we obtained. The methodology and further exploratory analysis of the same are detailed in this report.

Methodology



Beginning from data collection to model testing, this Hackathon was executed under semi-well defined steps as explain below:

1. Data collection: Our dataset consists of five excel (xlsx) files containing census data downloaded from the official Indian Censuses website. Data on migration, religion, occupations, languages and mother tongue for the year 2011 was collected.
2. Data cleaning and preprocessing: Since the datasets were processed according to the requirements of the collectors, we had to investigate all excel files, correct some column names, remove some unwanted columns and rows. Furthermore, we loaded each file into python and did the final cleaning. The following are some major steps taken;
   - *Removal of unwanted columns*: While exploring the excel files manually, we deleted unwanted columns like table name, state code, and district code which are not relevant for our analysis.
   - *Removal of unwanted rows:* The datasets downloaded contained aggregated rows as well as data for each state. Since we were interested in state wise data rather than country data, we deleted country data.
   - *Selection of aggregates:* Some of the data contained a total of certain categories as well as those categories. In areas where we needed the total, we collected only totals and discarded the others and in areas where we didn't need totals, we discarded all totals.

3. Calculation of Simpson's diversity index (SDI): It was initially designed to measure biodiversity but can be used to measure the diversity of places too. SDI ranges from 0 to 1 with diversity increasing as we approach one. It can be used to compare the diversity of two or more places.

$$D_{Simpson's} = 1 - \left(\Sigma \frac{n(n-1)}{N(N-1)}\right)$$

Where n is the number of instances in one class and N is the total number of instances across all classes. Using this tool, we computed the migration, religion, occupation and mother tongue diversity indices.

4. Calculation of Shannon's diversity index: This was also designed to measure biodiversity.

$$H_{Shannon} = -\sum_{i=1}^{n} p_i \ln\ p_i$$

Where p_i is the proportion of the ith species in the community.

5. Calculation of cultural diversity index: The cultural diversity index was calculated by taking the average of the migration, occupation, religion, and language indices. This was done for both Shannon and Simpson's indices for migration, occupation, religion and language.

$$CD = \frac{Sum\ of\ diversity\ index\ values}{Total\ number\ of\ indices}$$

6. Modelling: After calculations of all indices, we tried to model the final result using a linear model but since the response variable was an average of the exploratory variables, the model picked up on that and produced an accuracy of 100%, so we discarded the model. We used PCA to reduce the 4 indices down to two principal components. Which we then plotted and clustered using the K-Means algorithm with the approximated optimal cluster size; but since we could not understand the output of the PCA model, we dropped PCA.

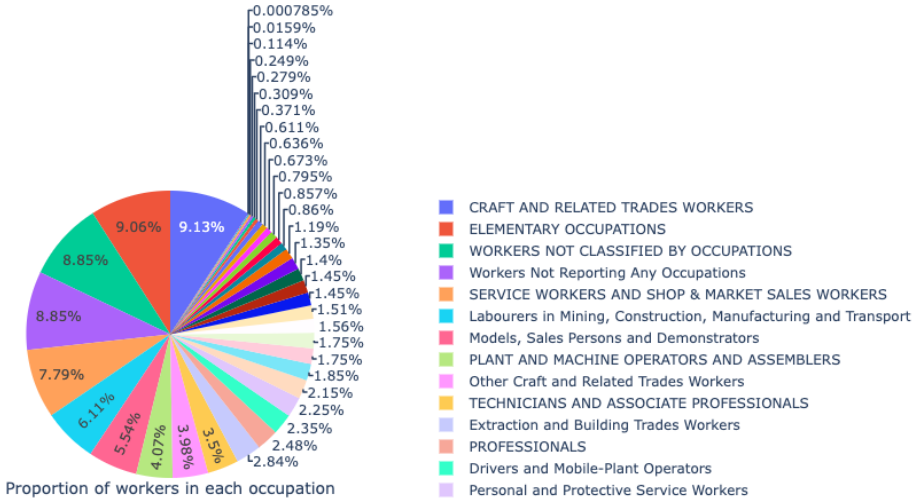
Exploratory Data Analysis

We conducted an EDA on the 2011 Indian Census Data in order to understand the data further:

Proportion of workers in each occupation

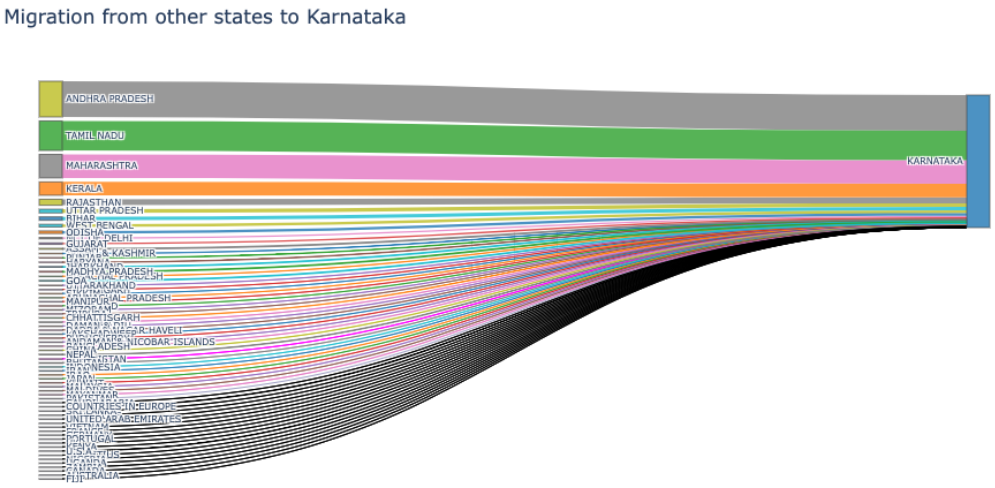
Different occupation classes of main and marginalised workers in India.

We note that craft and related trade workers have a higher proportion when compared with other occupations.



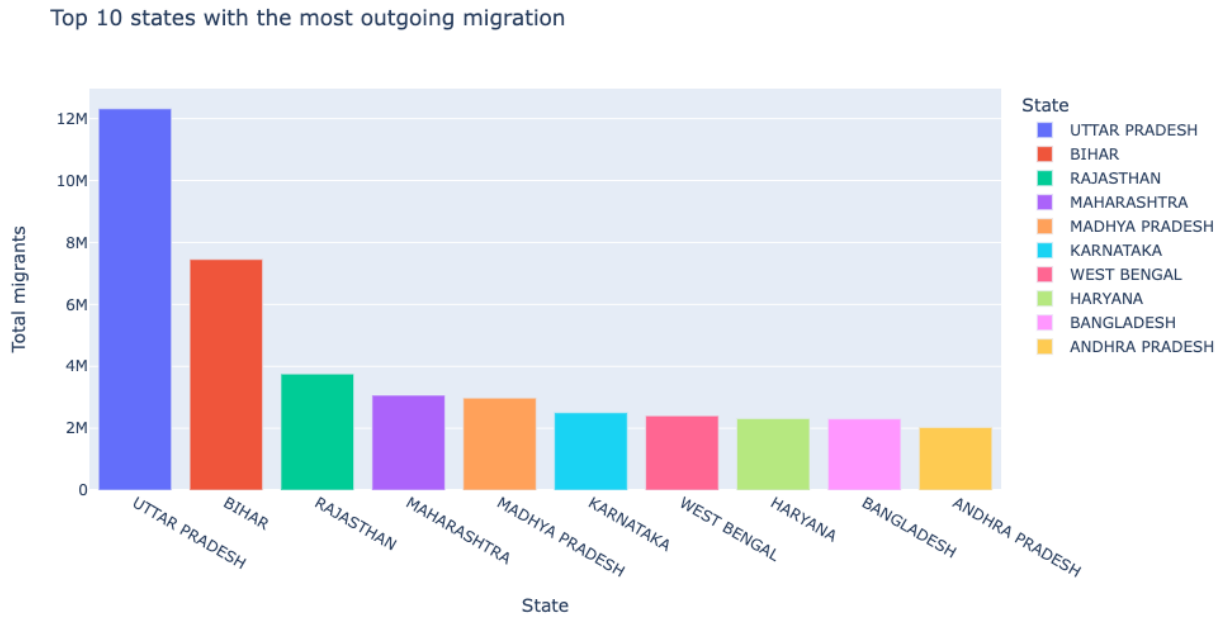
Migration from other states to Karnataka

We can see below that the people from Andhra Pradesh have migrated the most to Karnataka when compared to Migration from other states. Then comes Tamil Nadu followed by Maharashtra and Kerala. And the least comes from Fiji.



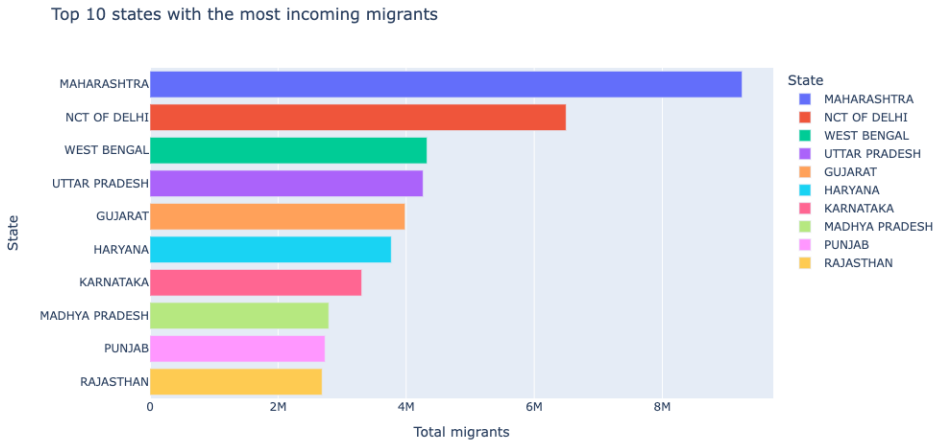
Top 10 states with the most outgoing migration

Here, we see that people from Uttar Pradesh migrate to other places the most, we assume that this could be because they seek opportunities out of state. The next states are Bihar and then Rajasthan with Andhra Pradesh being 10th in migrating out of state.



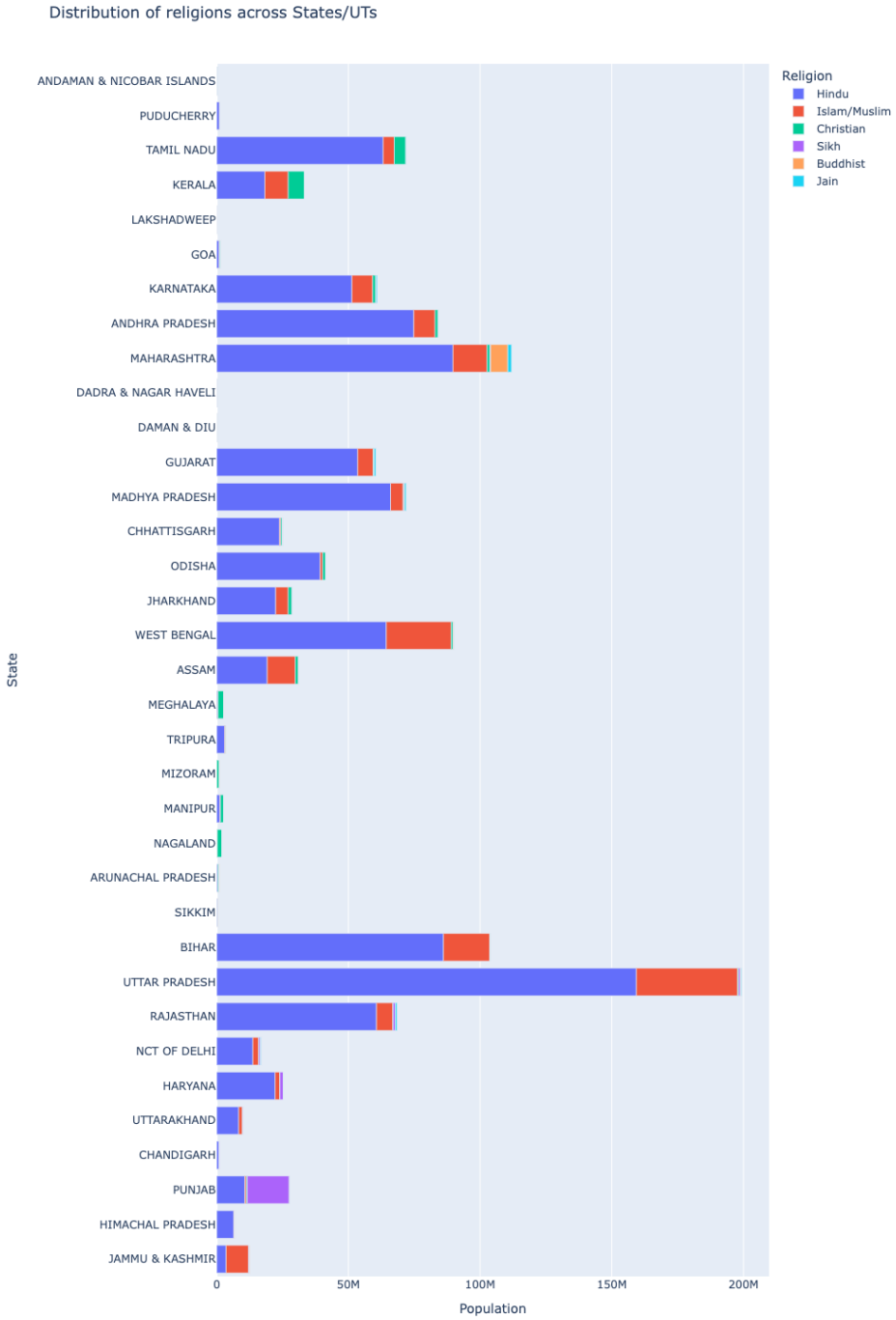
Top 10 states with the most incoming migrants

Of the states with most incoming migrants, Maharashtra stands first. People go to Maharashtra to find the opportunities for better standards of living. The next place people go to is the National Capital Territory of Delhi. With a little difference West Bengal and Uttar Pradesh stand third and fourth respectively. Rajasthan is in the tenth place with a lesser difference with Punjab which is in the Ninth place.



Distribution of religions across States/UTs

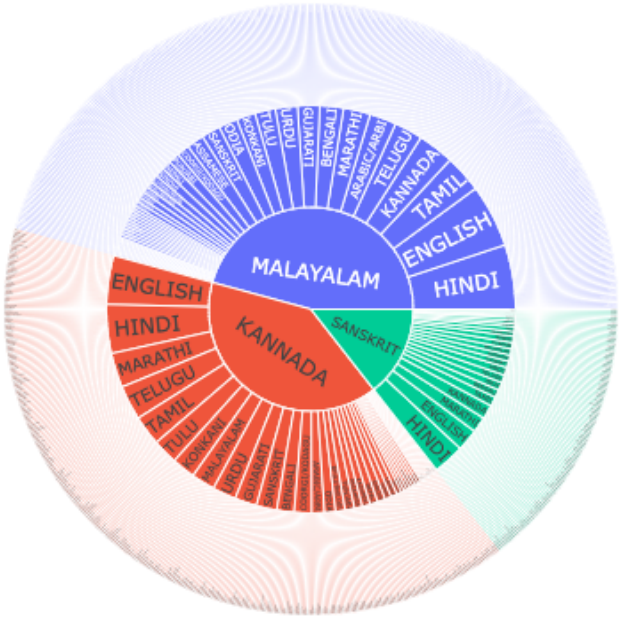
In the distribution of religions across states and union territories, we see that the majority of states have Hinduism as their prominent religious group with Uttar Pradesh having the highest density of Hindus. Followed by Maharashtra.



Proportion of first language speakers who can also speak a secondary and tertiary langauge

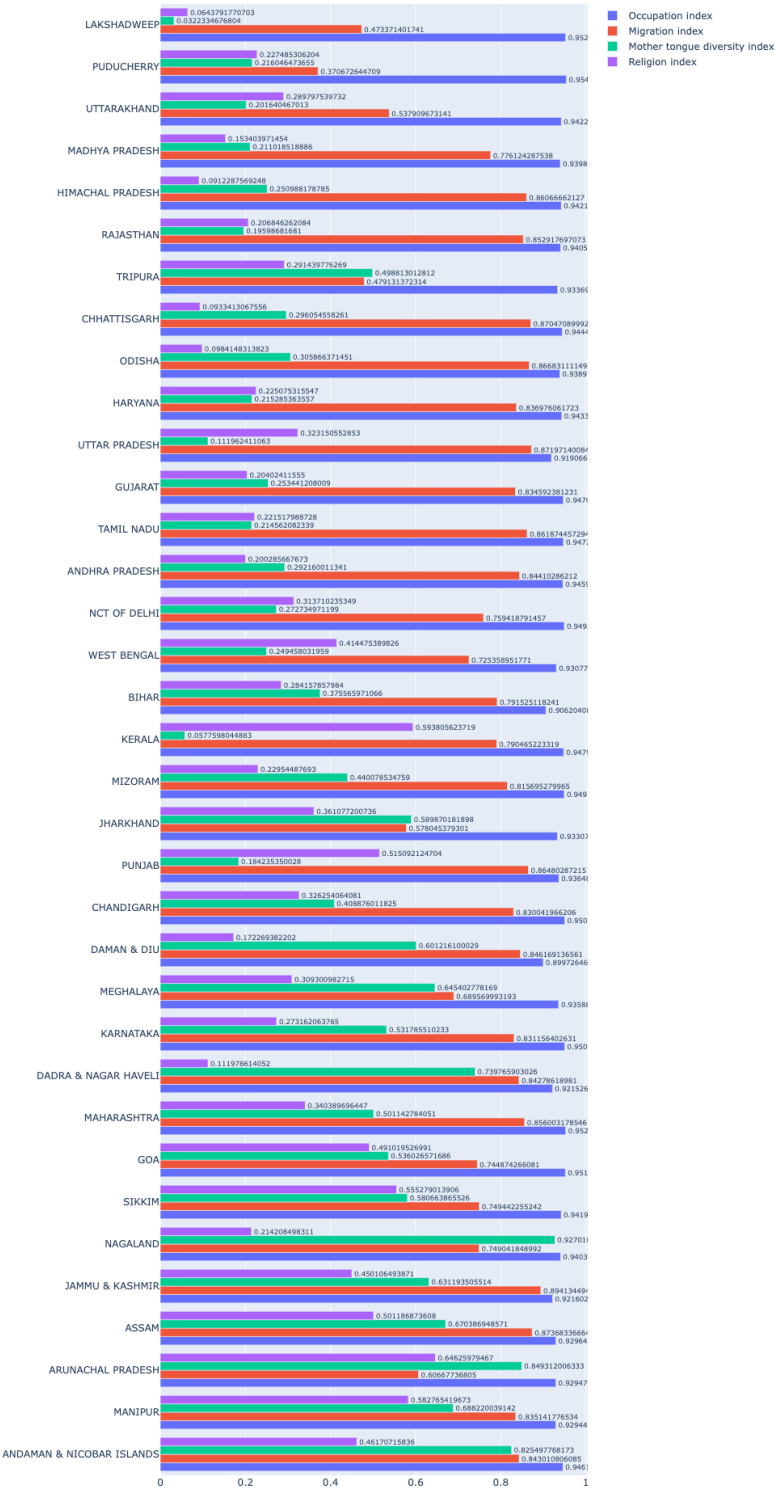
In the sunburst chart below, the first ring beginning from the innermost ring, represents the proportion of people who have Kannada, Sanskrit and Malayalam as their first language. Within the second ring, the languages in blue represent the proportion of people who speak their first language and a secondary language like Hindi. For each of the languages in the second ring, we noticed that there are some languages in the third ring that indicate a third language that is spoken by that same group of people.

The sunburst chart shows that the proportion of people who speak Malayalam as their first language is almost similar to that of those who speak Kannada as their first language, with Sanskrit being the least spoken. Furthermore, quite a large number of people who speak Malayalam as their first language also speak Hindi.



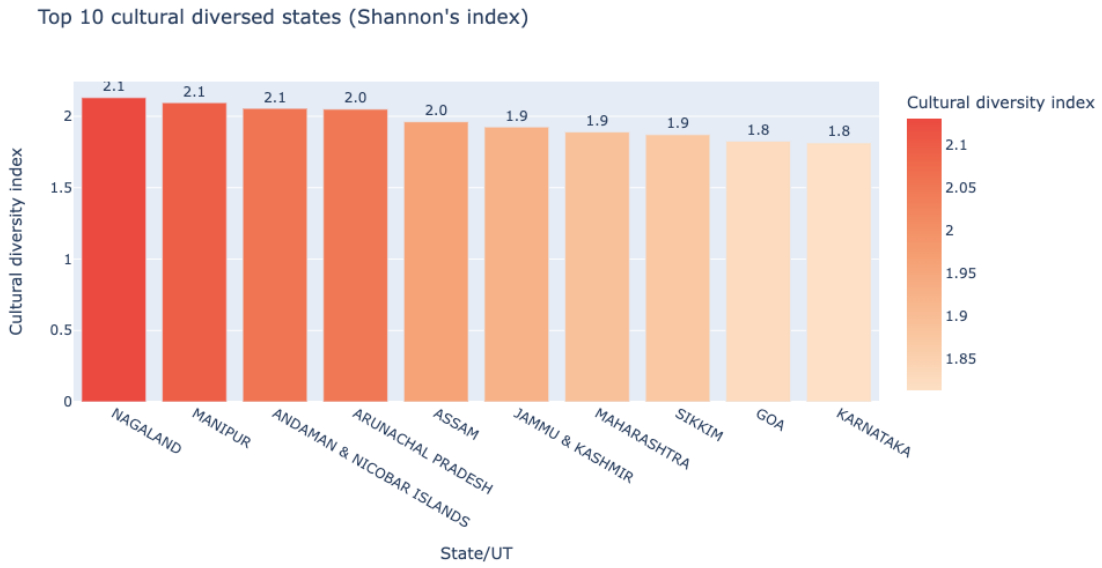
The four different indices for each state

The following chart shows the occupational diversity index, migration diversity index, linguistic diversity index and religious diversity index for each state.



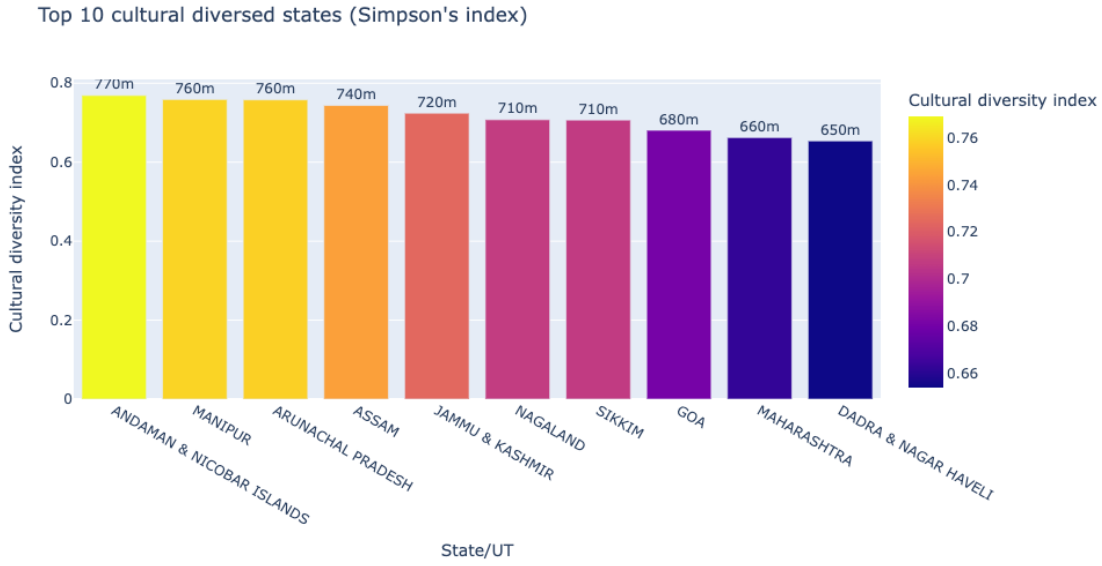
Top 10 cultural diverse states (Shannon's index)

In the top 10 culturally diverse states which was calculated based on the Shannon's diversity index of each aspect shows that Nagaland stands first followed by Manipur. And in 10th position is Karnataka.



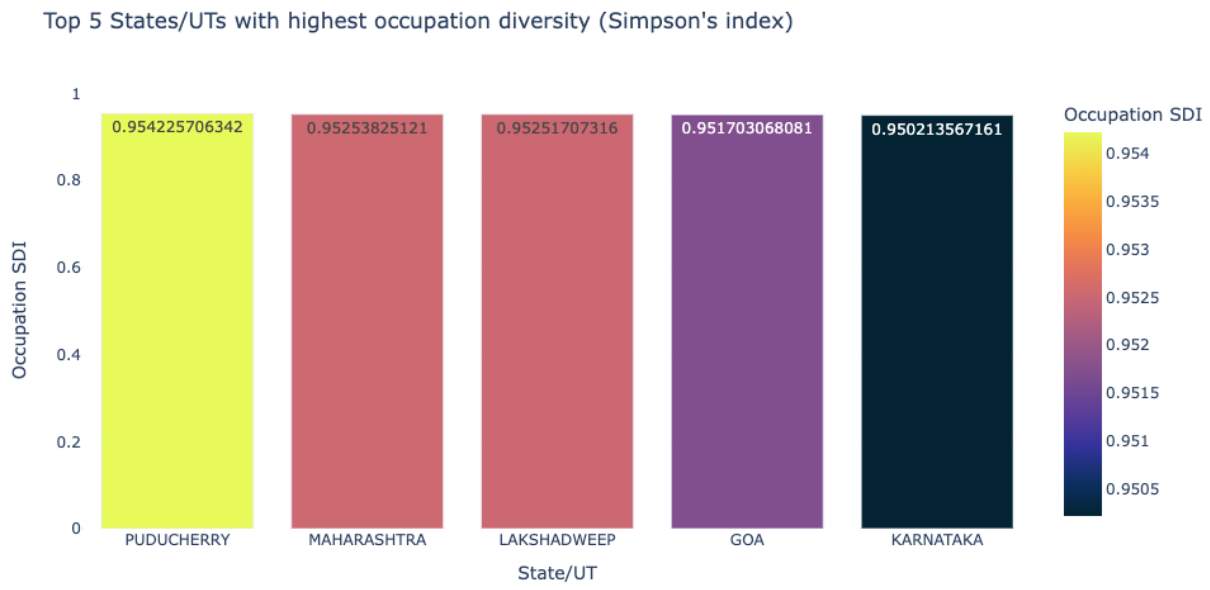
Top 10 cultural diverse states (Simpson's index)

In the top 10 culturally diverse states which was calculated based on Simpson's diversity index, Andaman & Nicobar Islands stands first followed by Manipur and in 10th position is a union territory named Dadra & Nagar Haveli.



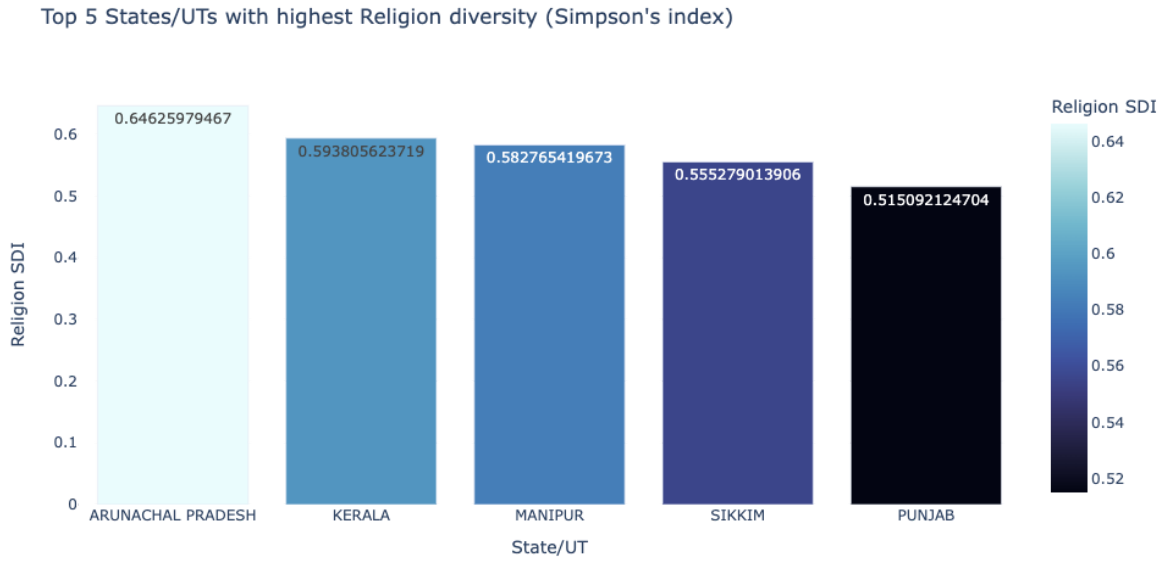
Top five states with high occupational diversity scores

Here we could see that Puducherry stands first in the top 5 in the highest occupation diversity calculated using Simpson's Index. And in fifth place is Karnataka.



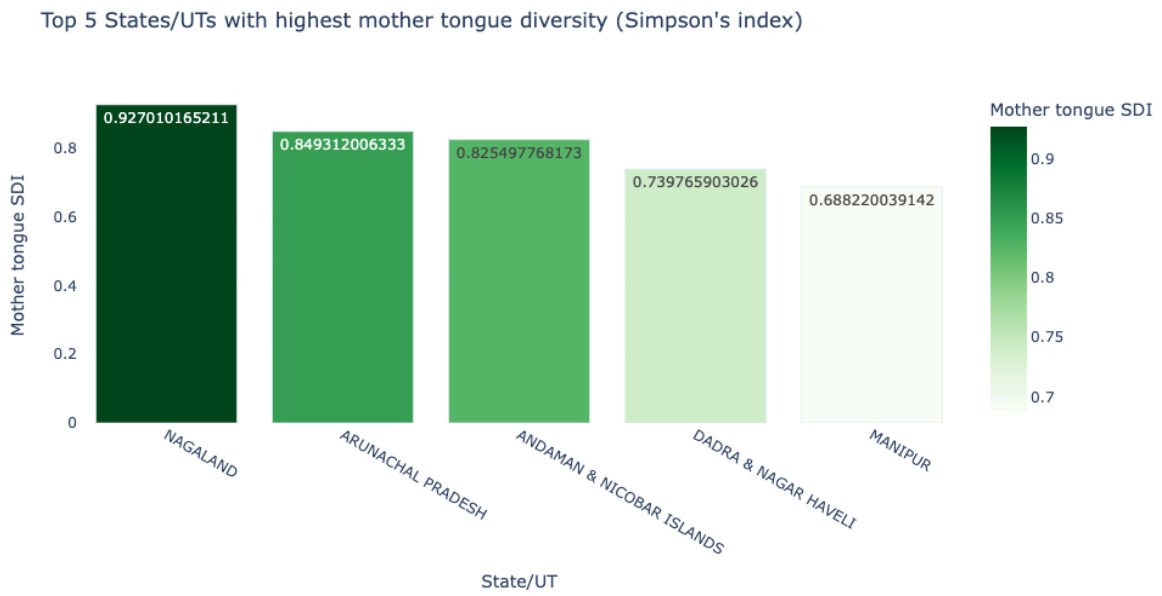
Top five states with high religious diversity scores

While checking for the highest Religion diversity calculated using the Simpson's Index, we can see that Arunachal Pradesh in the first and Punjab in the fifth position.



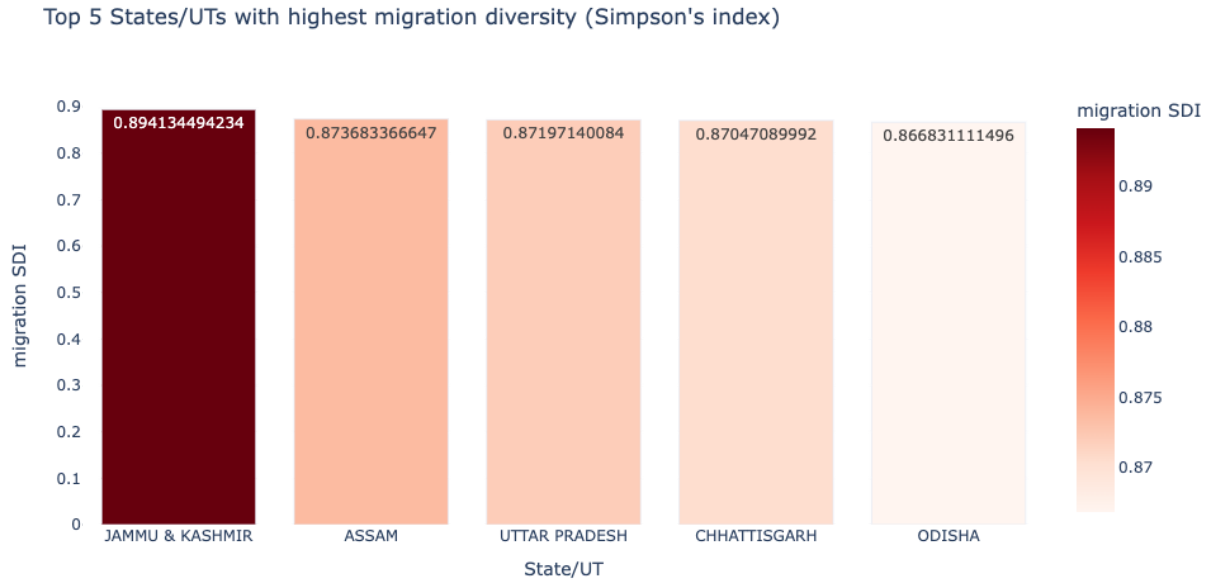
Top five states with high religious diversity scores

When coming to the highest mother tongue diversity, we can see Nagaland with the most diverse state for mother tongues. And in fifth position is Manipur.



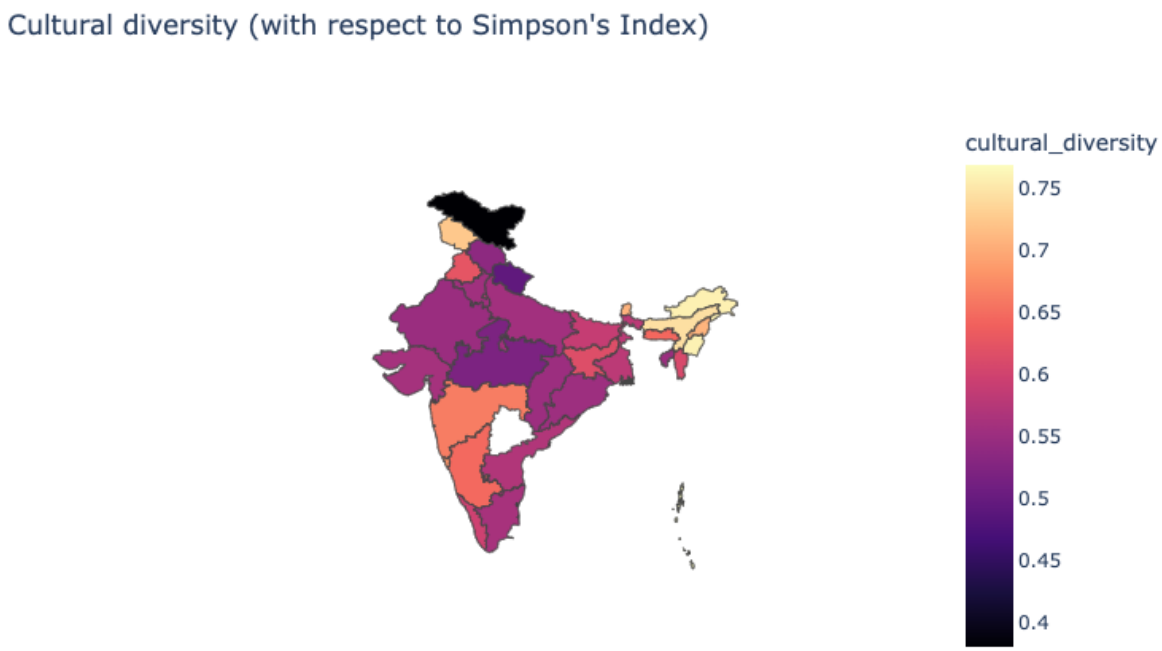
Top five states with high religious diversity scores

Here, we can see that Jammu & Kashmir has the most diversity in the migration. In the fifth position comes Odisha.



Cultural diversity of each state as represented using the heatmap on the map of India

The map of India above is a heatmap representation of the cultural diversity (with respect to Simpson's index) of each State/UT. The heatmap ranges from black representing low cultural diversity to light yellow representing a high cultural diversity. We can observe Ladakh (top south of India) has a very low cultural diversity whereas Arunachal Pradesh in the east has a high cultural diversity.



Conclusion & Inferences

In this Hackathon, using the 2011 census, we analysed the cultural diversity of each state in India. Using the Simpson and Shannon's diversity index, we calculated the migration, occupation, language and religion indices of each state. We later took the average of the aforementioned indices to serve as the cultural index of each state. Moreso, by first language, we analysed the diversity of second and third languages spoken by the same group of people. From our analysis, we deduced that Hinduism is the majority religion in many states of india. States like Manipur, Assam, Goa and Sikkim appeared in the top 10 states with highest cultural diversity for both Simpson and Shannon's indices. In addition, it was found that the highest migration, first language, religion and occupation diversities were found in Jammu and Kashmir, Nagaland, Arunachal Pradesh and Puducherry, respectively.

This information can help us better understand the people who come from various states and the wealth of knowledge and culture that they carry. This is critical in many decision making processes for policy makers, businesses and even individuals who have a desire to migrate to another state or union territory in India.

Future work

There is little work on a formula for cultural diversity among our human civilization. We'd like to uncover what models and approaches might be best when handling highly sensitive and quantitative in nature

- We did attempt to apply a linear model to find the linear combination of the four indices we calculated and their contribution towards an unknown overall effect
- PCA was also implemented in the attempt to uncover any underlying component that picks up most of the variation from the attributes

Key learning

This hackathon served as an opportunity for us to browse through census data and see how the government structures the data. In the beginning of the hackathon, it was difficult for us to get a good problem statement as the dataset was structured according to the needs of those who collected the data (government). However, with much thought we agreed to work on cultural diversity. Through this, we learned about the Simpson and Shannon indices for measuring biodiversity (which has been used on cultural diversity too) and we learned that there is no widely accepted method for measuring cultural diversity. Furthermore, cleaning the dataset helps us increase our experience with transforming data from raw form to a more useful form (as per our requirements). Lastly, we learned about different visualisation charts, and the package “plotly” (a visualisation library built on Python).

References

1. Simpson's Diversity Index (Simpson's Diversity Index: Definition, Formula, Calculation - Statistics How To)
2. Measuring cultural diversity with the Stirling model
3. Measuring a new aspect of ethnicity: The appropriate diversity index
4. Biolinguistic Diversity Index of India
5. Measuring the Diversity of Cultural Expressions: Applying the Stirling Model of Diversity in Culture
6. Measuring Racial and Ethnic Diversity for the 2020 Census

Start Date: *August 12th, 2023*.

End Date: *August 19th, 2023*.