CSC 324: Project Documentation

1. **Basic Documentation**
   1. **Purpose**

By enabling us to communicate with one another, language is what truly enables social life and helps us build communities. Through its use across thousands of years, any language is a corpus of collective cultural and historical knowledge, and it is imperative to study languages to understand what we gain from them, as well as what we stand to lose in the rapidly homogenizing global linguistic landscape.

This app uses India as a case study, due to its immense linguistic diversity, and serves as a starter pack visualisation for the languages of India. It is aimed at anyone with even a fledgling interest in the languages of the country, regardless of their level of knowledge of the region, and aims to answer questions relating to the languages and language families in the region, endangered languages, and the geographic distribution of all three of the aforementioned aspects.

* 1. **Data description and method of data collection**
     1. **Census 2011 Data**

Source Link: <https://censusindia.gov.in/2011census/C-16.html>

This dataset is the 2011 Population Mother Tongue Table taken from the Indian government’s census site. It consists of 10,333 observations and 15 columns – namely, Table, state\_code, district\_code, subdistrict\_code, area\_name, language\_code, language\_name, total, male, female, total\_rural, male\_rural, female\_rural, total\_urban, male\_urban, and female\_urban.

According to the Indian government’s census website, the data is collected by government workers who visit each house to collect the information and fill up 2 forms per household with their observations.

* + 1. **Languages of the world data**

Source Link: <https://www.kaggle.com/rtatman/world-atlas-of-language-structures>

The data was taken from the WALS (World Atlas of Language Structure) database, which is an online database maintained by the Max Planck Institute for Evolutionary Anthropology. It is actually a phonetic inventory of every language in this database, so it has 202 columns. The ones relevant to this project are name, latitude, longitude, family, and countrycode.

* + 1. **Endangered Languages data**

Source Link: <https://www.kaggle.com/the-guardian/extinct-languages>

The dataset was taken from a blogpost on Guardian. The relevant columns are the language name in English, Number of speakers of that language, latitude, longitude and degree of endangerment.

* + 1. **Shapefile**

Source Link: <https://www.diva-gis.org/gdata>

This data was provided by GADM. The specific shapefile I used is a shapefile that captures the geometric specifications of the states of India.

* 1. **Questions and strengths**

Questions:

* What languages exist in India and what language families do they belong to?
* What is the geographic distribution of these language families?
* What languages are endangered in India and what is the status of it?
* How are endangered languages distributed?
* Is there a correlation between number of languages in a family and the number of speakers of that language
* What is the distribution of the speakers of each of the 20 scheduled languages across the states of India?
  1. **Insights from data**
* "Some insights I have noticed by creating this app are that the Sino-Tibetan languages seem to be spread out along the border between India and China. The Dravidian languages seem to span the southern states of India, while Austro-Asian languages seem to occur to the east of India. Indo-European languages span across the country but it the only language family, with languages that occur in the west of India. Finally, the north and south Andamese isolates solely occur on the Andaman and Nicobar islands.
* The Sino-Tibetan Language family has the greatest linguistic diversity, with it having 54 languages in it.
* The Indo-European language family has the most number of speakers, despite being the third most linguistically diverse family, and has about a billion speakers
* Linguistics diversity of a family is not necessarily predictive of number of speakers of the languages in that family
* Endangered languages seem to occur most frequently along the borders of the country and in regions where there seems to be a confluence of many language families or many languages. This may be due to increased competition between the increased number of languages.
* Hindi is, by far, the most spoken language in the country and even one 22 the official languages of India, Bodo, is an endangered language.

* 1. **Possible improvements and extensions**

I would change the maps in the Endangered Languages section of the app so that the size of each circle marker corresponds to the number of speakers of that language. This would allow for all the information on the labels to be visually recognizable.

An interesting idea that struck me while creating and looking at the various maps was how the distribution of endangered languages seems to coincide with areas of greater linguistic density, in that many languages are spoken there, as well as regions where there is a confluence of various different language families. Leading from that, the endangered languages section of this app could be extended to find out the factors that may be predictive of a greater number of endangered languages occurring in a geographic region. This section could use both maps as well as scatterplots to shows the cooccurrence of endangered languages with different factors, as well as the correlation of factor and the occurrence of endangered languages.

I would also like to make the geographic distribution of the speaker population density for a selected language to have a slider for different years in accordance with the census from previous and future years. This allows us to see how the number of speakers of a particular language, as well as their geographic distribution in India, has changed over time.

It would also be fascinating to investigate whether the linguistic diversity of a language family has any correlation with the number of speakers of languages in the family through making a scatterplot of the two variables.

1. **Describing the process and development to make your work reproducible, for example, *tidying data*.**

All geographic plots were made using the leaflet package, while all pie and doughnut charts were made using the plotly package.

In order to create the ‘Geographic distribution of Indian language families’ map, I used the Languages of the world data. This data set was clean and did not require any tidying. I simply plotted each of the data points on the map using the latitude and longitude variables and made it so the circle markers we assigned colours in accordance with the language family the language belonged to. I also used this data to make the ‘Number of languages in language families’ doughnut chart.

In making the ‘Number of total speakers per language family’ doughnut chart, I used the aforementioned data, and merged it with the Census 2011 data to then be able to get the total number of speakers per language family. The census data is extremely messy. The first step is to rename the columns of the data set as the original column names are very long, span over two rows, and are not very descriptive. Then, I would remove any white space using trimws(<df$col>, which=”both”) to ensure any future joins or merges would work. Any columns that need to be used for the joins or merges should also have a standardized name across the two tables.

To make all of the maps in the ‘Mapping endangered languages’ section, I used the Endagered Languages dataset. This data set is also extremely clean, so all I did was plot each of the points with the latitude and longitude, and then assigned the circle markers color in accordance with their degree of endangerment.

The ‘Geographic distribution of speaker population density of a Selected Langauge’ maps were all made using the Census data as well as the Shapefile. The relevant data for each of the 20 languages was extracted using the country code, the name of language and states was changed to fit the format of those in the shapefile, and the column for the language was merged with the shapefile using the name of the states. Any NA values were replaced with 0, as the language has no speakers in the given state.

The ‘Scheduled languages of India with number of speakers’ map was created using the Census data, simply by considering the number of speakers for each language such that the area\_name of in the census data was “INDIA”.

1. **Description of design decisions (encoding/mapping). Use the taxonomy of the what-why-how analysis framework presented by Tamara Munzner.**

|  |  |  |  |
| --- | --- | --- | --- |
| Chart | What? | Why? | How? |
| Geographic distribution of Indian Language Families Map  Map  Description automatically generated | The spatial mapping of the latitude and longitude associated with each language in India. Additionally, categorical attributes coded through the colour. | Users can use this map to discover the languages of India, their geographic location, as well the trends in geographic distribution of the different language families in the country. | Color-coded circle markers denoting a language mapped on a base map. The colour corresponds to the language family of the language. |
| Chart, pie chart  Description automatically generated | A table consisting of one quantitative attribute and one (number of languages in the a language family) categorical attribute (language family). | Look-up values, allows users to understand the make-up of the language families of India | A polar areal chart whose coloured sections corresponds to different languages families. The area of each section corresponds to the number of languages in that family. |
| Chart, pie chart  Description automatically generated | A table consisting of one quantitative attribute (total speakers) and one categorical attribute(language family). | Look-up values, allows users to understand the make-up of the total speakers of language families of India. | A polar areal chart whose coloured sections corresponds to different languages families. The area of each section corresponds to the number of total speakers of languages in the given family. |
| Map of endangered languages with degree of endangerment  Map  Description automatically generated | The spatial mapping of the latitude and longitude associated with each endangered  language in India. Additionally, categorical attributes coded through the colour. | Users can use this map to discover the endangered languages of India, their geographic location, as well the trends in geographic distribution of endangered languages. | Color-coded circle markers denoting a language mapped on a base map. The colour corresponds to the degree endangerment of the language. The labels give further information about the language (number of speakers). |
| Geographis distribution of the speaker population density of a selected language map  Map  Description automatically generated | Geographic geometric data (from the shapefile), along with the associated numeric field (number of speakers of selected language in that state) | Look up speakers of a language by state, view geographic distribution of speakers and observe possible trends in them. | The shapefile is used to denote the geometry of the state of India. The colour used is a sequential segmented colour map. The darker the color in a given state, the more speakers of the selected language it has. |
| Chart, pie chart  Description automatically generated | A table consisting of one quantitative attribute (total speakers) and one categorical attribute(language name). | Look-up values, allows users to understand the make-up of the total speakers of the 20 official languages of India. | A polar areal chart whose coloured sections corresponds to different official languages of India. The area of each section corresponds to the number of total speakers of that language. |

1. **Sources or References**

[1] Government of India Ministry of Home Affairs Office of the Registrar General and Census Comissioner, 2011. “Population by Mother Tongue,” distributed by Census India Government, <https://censusindia.gov.in/2011Census/Language_MTs.html>

[2] R. Tatman, 7th September, 2017. “World Atlas of Language Structures,” distributed by Kaggle, <https://www.kaggle.com/rtatman/world-atlas-of-language-structures>

[3] “Download Data by country,” *DIVA*. [Online]. Available: https://www.diva-gis.org/gdata. [Accessed: 14-Feb-2022].

[4] “Application layout guide,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/articles/layout-guide.html. [Accessed: 14-Mar-2022].

[5] “Download Data by country,” *DIVA*. [Online]. Available: https://www.diva-gis.org/gdata. [Accessed: 14-Mar-2022].

[6] “Lesson 3 add control widgets,” *Shiny*. [Online]. Available: https://shiny.rstudio.com/tutorial/written-tutorial/lesson3/. [Accessed: 14-Mar-2022].

[7] NaibafNaibaf 37744 silver badges1414 bronze badges, timelyportfoliotimelyportfolio 6, and conrad-macconrad-mac 78177 silver badges1212 bronze badges, “Can't loop with R's leaflet package to produce multiple maps,” *Stack Overflow*, 01-Feb-1964. [Online]. Available: https://stackoverflow.com/questions/36554605/cant-loop-with-rs-leaflet-package-to-produce-multiple-maps/36587525#36587525. [Accessed: 14-Mar-2022].

[8] National Geographic Society, “Family of language,” *National Geographic Society*, 17-Jul-2020. [Online]. Available: https://www.nationalgeographic.org/encyclopedia/family-language/. [Accessed: 14-Mar-2022].

[9] P. by P. Chan, “Researcher@Library blog,” *ResearcherLibrary Blog*, 19-Jun-2020. [Online]. Available: https://blogs.unimelb.edu.au/researcher-library/2020/06/19/making-a-covid-19-map-in-r-using-shiny-and-leaflet/. [Accessed: 14-Mar-2022].

[10] Rstudio, “Leaflet/labels.r at main · RSTUDIO/Leaflet,” *GitHub*. [Online]. Available: https://github.com/rstudio/leaflet/blob/main/inst/examples/labels.R. [Accessed: 14-Mar-2022].

[11] S. Jolad and A. Agarwal, “India's linguistic diversity,” *The India Forum*, 19-Dec-2021. [Online]. Available: https://www.theindiaforum.in/article/what-census-obscures. [Accessed: 14-Mar-2022].

[12] T. Guardian, “Extinct languages,” *Kaggle*, 07-Dec-2016. [Online]. Available: https://www.kaggle.com/the-guardian/extinct-languages. [Accessed: 13-Mar-2022].

[13] UNESCO, *Atlas of the world's Languages in danger: 3rd revised edition*. United Nations Educational, Scientific & Cult, 2010.

[14] VP Nagraj, “Choropleth maps with leaflet,” *VP Nagraj*, 25-Jul-2016. [Online]. Available: https://www.nagraj.net/notes/choropleth-maps-with-leaflet/. [Accessed: 14-Mar-2022].

[15] Y. Holtz, “Interactive choropleth map with R and leaflet,” *– the R Graph Gallery*. [Online]. Available: https://www.r-graph-gallery.com/183-choropleth-map-with-leaflet.html. [Accessed: 14-Mar-2022].

[16] “[R shiny basic app] #7 interactive ggplot bar ... - youtube.” [Online]. Available: https://www.youtube.com/watch?v=rlcUiSw49FY. [Accessed: 14-Mar-2022].