India is the world’s largest Democracy and as it goes, also a highly diverse place. This is my attempt to see how “Hindi” and other languages are spoken in India.

In this post, we’ll see how to collect data for this relevant puzzle – directly from Wikipedia and How we’re going to visualize it – highlighting the insight.

**Data**

Wikipedia is a great source for data like this – Languages spoken in India and also because Wikipedia lists these tables as html

|  |
| --- |
|  |

it becomes quite easier for us to use rvest::html\_table() to extract the table as dataframe without much hassle.

options(scipen = 999)

library(rvest) # for webscraping

library(tidyverse) # for data analysis and visualization

# the wikipedia page URL - thanks to DuckDuckGo search

lang\_url <- "<https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India>"

# extracting the entire content of the page

content <- read\_html(lang\_url)

# extracting only tables from the downloaded content

tables <- content %>% html\_table(fill = TRUE)

# from the page we know, it's the first table we want picking up the first

# element from the list of tables

lang\_table <- tables[[1]]

### header cleaning - exclude the first row

lang\_table <- lang\_table[-1, ]

lang\_table

## First language speakers First language speakers

## 2 Hindi[b] 422,048,642 41.03%

## 3 English 226,449 0.02%

## 4 Bengali 83,369,769 8.10%

## 5 Telugu 74,002,856 7.19%

## 6 Marathi 71,936,894 6.99%

## 7 Tamil 60,793,814 5.91%

## 8 Urdu 51,536,111 5.01%

## 9 Kannada 37,924,011 3.69%

## 10 Gujarati 46,091,617 4.48%

## 11 Odia 33,017,446 3.21%

## 12 Malayalam 33,066,392 3.21%

## 13 Sanskrit 14,135 <0.01%

## Second languagespeakers[11] Third languagespeakers[11] Total speakers

## 2 98,207,180 31,160,696 551,416,518

## 3 86,125,221 38,993,066 125,344,736

## 4 6,637,222 1,108,088 91,115,079

## 5 9,723,626 1,266,019 84,992,501

## 6 9,546,414 2,701,498 84,184,806

## 7 4,992,253 956,335 66,742,402

## 8 6,535,489 1,007,912 59,079,512

## 9 11,455,287 1,396,428 50,775,726

## 10 3,476,355 703,989 50,271,961

## 11 3,272,151 319,525 36,609,122

## 12 499,188 195,885 33,761,465

## 13 1,234,931 3,742,223 4,991,289

## Total speakers

## 2 53.60%

## 3 12.18%

## 4 8.86%

## 5 8.26%

## 6 8.18%

## 7 6.49%

## 8 5.74%

## 9 4.94%

## 10 4.89%

## 11 3.56%

## 12 3.28%

## 13 0.49%

At this point, we’ve got the required table but mind you, The numbers are in characters and for us to plot visualizations – it has to be in Numeric format. We’ll pick only **First Language Speakers** for further sections so will change those numbers from character into numeric format

# clean-up the messed up column names

lang\_table <- lang\_table %>%

janitor::clean\_names()

lang\_table[1,"x"] <- "Hindi"

lang\_table

## x first\_language\_speakers first\_language\_speakers\_2

## 2 Hindi 422,048,642 41.03%

## 3 English 226,449 0.02%

## 4 Bengali 83,369,769 8.10%

## 5 Telugu 74,002,856 7.19%

## 6 Marathi 71,936,894 6.99%

## 7 Tamil 60,793,814 5.91%

## 8 Urdu 51,536,111 5.01%

## 9 Kannada 37,924,011 3.69%

## 10 Gujarati 46,091,617 4.48%

## 11 Odia 33,017,446 3.21%

## 12 Malayalam 33,066,392 3.21%

## 13 Sanskrit 14,135 <0.01%

## second\_languagespeakers\_11 third\_languagespeakers\_11 total\_speakers

## 2 98,207,180 31,160,696 551,416,518

## 3 86,125,221 38,993,066 125,344,736

## 4 6,637,222 1,108,088 91,115,079

## 5 9,723,626 1,266,019 84,992,501

## 6 9,546,414 2,701,498 84,184,806

## 7 4,992,253 956,335 66,742,402

## 8 6,535,489 1,007,912 59,079,512

## 9 11,455,287 1,396,428 50,775,726

## 10 3,476,355 703,989 50,271,961

## 11 3,272,151 319,525 36,609,122

## 12 499,188 195,885 33,761,465

## 13 1,234,931 3,742,223 4,991,289

## total\_speakers\_2

## 2 53.60%

## 3 12.18%

## 4 8.86%

## 5 8.26%

## 6 8.18%

## 7 6.49%

## 8 5.74%

## 9 4.94%

## 10 4.89%

## 11 3.56%

## 12 3.28%

## 13 0.49%

lang\_table %>%

select(one\_of("x","first\_language\_speakers")) %>%

mutate(first\_language\_speakers = parse\_number(first\_language\_speakers)) -> lang\_table\_first

names(lang\_table\_first) <- c("Language","first\_language\_speakers")

lang\_table\_first

## Language first\_language\_speakers

## 1 Hindi 422048642

## 2 English 226449

## 3 Bengali 83369769

## 4 Telugu 74002856

## 5 Marathi 71936894

## 6 Tamil 60793814

## 7 Urdu 51536111

## 8 Kannada 37924011

## 9 Gujarati 46091617

## 10 Odia 33017446

## 11 Malayalam 33066392

## 12 Sanskrit 14135

**Visualization**

Now that we got a categorical and a numerical variable. It’s time to play with some visualization – as it’s typical – a bar chart.

**All Languages**

lang\_table\_first %>%

mutate(Language = fct\_reorder(Language,-first\_language\_speakers)) %>%

ggplot() + geom\_bar(aes(Language, first\_language\_speakers),

stat = "identity",

fill = ifelse(lang\_table\_first$Language == 'Hindi',

"#ffdd00",

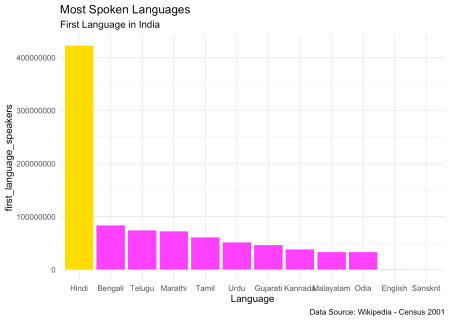
"#ff00ff")) +

theme\_minimal() +

labs(title = "Most Spoken Languages",

subtitle = "First Language in India",

caption = "Data Source: Wikipedia - Census 2001")



That’s a long tail with Hindi leading the way.

**Hindi & Everyone else**

library(viridis)

lang\_table\_first %>%

mutate(Language = ifelse(Language == "Hindi",

"Hindi","non\_Hindi")) %>%

group\_by(Language) %>%

summarize(first\_language\_speakers = sum(first\_language\_speakers)) %>%

mutate(percentage = round((first\_language\_speakers / sum(first\_language\_speakers))\*100,2)) %>%

ggplot() + geom\_bar(aes(Language,percentage,fill = Language), stat = "identity"

) +

scale\_fill\_viridis\_d(option = 'E', direction = -1) +

scale\_y\_continuous(limits = c(0,60)) +

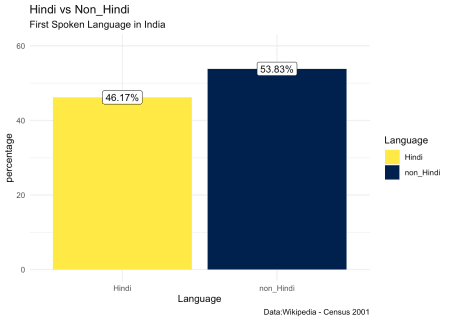
theme\_minimal() +

geom\_label(aes(Language,percentage, label= paste0(percentage,"%"))) +

labs(title = "Hindi vs Non\_Hindi",

subtitle = "First Spoken Language in India",

caption = "Data:Wikipedia - Census 2001")



Living up to the Diversity of India, A mixed (assorted) group of languages other than Hindi forms ~54% while Hindi-only is ~46%