

Mahitha Valluru

Homework 7

DS4100: Data Science

## Creating a Data Frame of the Demographics of India – More Specifically the Population Distribution by States

Assignment 7 indicates that I must choose a web scraping toolkit to scrape the data from a website of my choice.

I chose Google Sheets as a web scraping toolkit, with its simple yet powerful ImportHtml() function. With this toolkit, I planned on scraping a Wikipedia page on the Demographics of India. On this page are many tables of useful information, but the one I chose is labeled “Population Distribution by states/union territories (2011).”

Here is a screenshot of the Wikipedia landing page:

The screenshot shows the Wikipedia article for "Demographics of India". The page includes a sidebar with navigation links, a main content area with text and a table of contents, and a right-hand box with a map and demographic statistics.

**Demographics of India**

From Wikipedia, the free encyclopedia

This article is about the people from India. For other uses, see *Indian (disambiguation)*.

India is the second most populated country in the world with nearly a fifth of the world's population. According to the 2017 revision of the World Population Prospects<sup>[1]</sup>, the population stood at 1,324,171,354. During 1975–2010 the population doubled to 1.2 billion. The Indian population reached the billion mark in 1998. India is projected to be the world's most populous country by 2022<sup>[1]</sup> surpassing the population of China. It is expected to become the first political entity in history to be home to more than 1.5 billion people by 2030, and its population is set to reach 1.7 billion by 2050.<sup>[2][3]</sup> Its population growth rate is 1.2%, ranking 94th in the world in 2013.<sup>[1]</sup>

India has more than 50% of its population below the age of 25 and more than 65% below the age of 35. It is expected that, in 2020, the average age of an Indian will be 29 years, compared to 37 for China and 48 for Japan, and, by 2030, India's dependency ratio should be just over 0.4.<sup>[4]</sup>

India has more than two thousand ethnic groups,<sup>[5]</sup> and every major religion is represented, as are four major families of languages (Indo-European, Dravidian, Austroasiatic and Sino-Tibetan languages) as well as two language isolates (the Nihali language<sup>[6]</sup> spoken in parts of Maharashtra and the Burushaski language spoken in parts of Jammu and Kashmir (Kashmir)). Further complexity is lent by the great variation that occurs across this population on social parameters such as income and education. Only the continent of Africa exceeds the linguistic, genetic and cultural diversity of the nation of India.<sup>[1][1]</sup>

The sex ratio is 944 females for 1000 males (2016).<sup>[12]</sup>

**Contents** (hide)

- History
  - 1.1 Prehistory to early 19th century
  - 1.2 Late 19th century to early 20th century
- Salient features
  - 2.1 Comparative demographics
  - 2.2 List of states and union territories by demographics
  - 2.3 Religious demographics
  - 2.4 Neonatal and infant demographics
  - 2.5 Population within the age group of 0–6
  - 2.6 Population above the age of 7
  - 2.7 Literacy rate
  - 2.8 Linguistic demographics
  - 2.9 Largest cities
- Vital statistics
  - 3.1 UN estimates
  - 3.2 Census of India: sample redistribution census

**Demographics of India**

Map showing the population density of each district in India.

<b>Population</b>	1,324,171,354 (2016 est.) <sup>[1]</sup>
<b>Density</b>	382 people per sq km (2011 est.)
<b>Growth rate</b>	▲ 1.19% (2016) (95th)
<b>Birth rate</b>	19.3 births/1,000 population (2016 est.)
<b>Death rate</b>	7.3 deaths/1,000 population (2016 est.)
<b>Life expectancy</b>	68.89 years (2009 est.)
<b>• male</b>	67.46 years (2009 est.)
<b>• female</b>	72.61 years (2009 est.)
<b>Fertility rate</b>	2.2 children born/woman (2016 est.) <sup>[1]</sup>
<b>Infant mortality rate</b>	41 deaths/1,000 live births (2016 est.) <sup>[12]</sup>

And here is the screenshot for the table that is of my interest:

Rank	State/UT	Population <sup>[40]</sup>	Percent (%)	Male	Female	Sex Ratio	Rural <sup>[40]</sup>	Urban <sup>[40]</sup>	Area <sup>[50]</sup> (km <sup>2</sup> )	Density (per km <sup>2</sup> )
1	Uttar Pradesh	199,812,341	16.50	104,480,510	95,331,831	930	155,111,022	44,470,455	240,928	828
2	Maharashtra	112,374,333	9.28	58,243,056	54,131,277	929	61,545,441	50,827,531	307,713	365
3	Bihar	104,099,452	8.60	54,278,157	49,821,295	918	92,075,028	11,729,609	94,163	1,102
4	West Bengal	91,276,115	7.54	46,809,027	44,467,088	950	62,213,676	29,134,060	88,752	1,030
5	Madhya Pradesh	72,626,809	6.00	37,612,306	35,014,503	931	52,537,899	20,059,666	308,245	236
6	Tamil Nadu	72,147,030	5.96	36,137,975	36,009,055	996	37,189,229	34,949,729	130,058	555
7	Rajasthan	68,548,437	5.66	35,550,997	32,997,440	928	51,540,236	17,080,776	342,239	201
8	Karnataka	61,095,297	5.05	30,966,657	30,128,640	973	37,552,529	23,578,175	191,791	319
9	Gujarat	60,439,692	4.99	31,491,260	28,948,432	919	34,670,817	25,712,811	196,024	308
10	Andhra Pradesh	49,386,799	4.08	24,738,068	24,648,731	996	34,776,389	14,610,410	160,205	308
11	Odisha	41,974,218	3.47	21,212,136	20,762,082	979	34,951,234	6,996,124	155,707	269
12	Telangana	35,193,978	2.91	17,704,078	17,489,900	988	21,585,313	13,608,665	114,840	307
13	Kerala	33,406,061	2.76	16,027,412	17,378,649	1084	17,445,506	15,932,171	38,863	859
14	Jharkhand	32,988,134	2.72	16,930,315	16,057,819	948	25,036,946	7,929,292	79,714	414
15	Assam	31,205,576	2.58	15,939,443	15,266,133	958	26,780,526	4,388,756	78,438	397
16	Punjab	27,743,338	2.29	14,639,465	13,103,873	895	17,316,800	10,387,436	50,362	550
17	Chhattisgarh	25,545,198	2.11	12,832,895	12,712,303	991	19,603,658	5,936,538	135,191	189
18	Haryana	25,351,462	2.09	13,494,734	11,856,728	879	16,531,493	8,821,588	44,212	573
19	Delhi (UT)	16,787,941	1.39	8,887,326	7,800,615	868	944,727	12,905,780	1,484	11,297
20	Jammu and Kashmir	12,541,302	1.04	6,640,662	5,900,640	889	9,134,820	3,414,106	222,236	56
21	Uttarakhand	10,086,292	0.83	5,137,773	4,948,519	963	7,025,583	3,091,169	53,483	189
22	Himachal Pradesh	6,864,602	0.57	3,481,873	3,382,729	972	6,167,805	688,704	55,673	123
23	Tripura	3,673,917	0.30	1,874,376	1,799,541	960	2,710,051	960,981	10,486	350
24	Meghalaya	2,966,889	0.25	1,491,832	1,475,057	989	2,368,971	595,036	22,429	132
25	Manipur	2,855,794	0.24	1,438,687	1,417,107	985	1,899,624	822,132	22,327	128
26	Nagaland	1,978,502	0.16	1,024,649	953,853	931	1,406,861	573,741	16,579	119
27	Goa	1,458,545	0.12	739,140	719,405	973	551,414	906,309	3,702	394
28	Arunachal Pradesh	1,383,727	0.11	713,912	669,815	938	1,069,165	313,446	83,743	17

As you can see, there is much information in this table and plenty of variables that I can (possibly in the future) analyze.

The column names include:

- State/Territory
- Population
- Percent population compared to the country as a whole
- Number of Males
- Number of Females
- Sex Ratio
- Population living in Urban Areas
- Population living in Rural Areas
- Area per square kilometer
- Density per square kilometer

When using the Google Sheets web scraping toolkit, I used the ImportHtml() function.

To use it, I needed to enter some information into the function in the top left-hand corner of the cells, so in cell A1:

1. URL of the page you would like to be scraped in quotes
2. Query which can either be a table or list in quotes
3. Index of the table you would like to be scraped as an integer

So, in this case, I typed:

```
=IMPORTHTML("https://en.wikipedia.org/wiki/Demographics_of_India", "table", 6)
```

And received this as a result:

Data Science - Demographics of India											
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive											
25712811											
	A	B	C	D	E	F	G	H	I	J	K
1	Rank	State/UT	Population[48]	Percent (%)	Male	Female	Sex Ratio	Rural[49]	Urban[49]	Area[50] (km2)	Density (per km2)
2	1	Uttar Pradesh	199,812,341	16.5	104,480,510	95,331,831	930	155,111,022	44,470,455	240,928	828
3	2	Maharashtra	112,374,333	9.28	58,243,056	54,131,277	929	61,545,441	50,827,531	307,713	365
4	3	Bihar	104,099,452	8.6	54,278,157	49,821,295	918	92,075,028	11,729,609	94,163	1,102
5	4	West Bengal	91,276,115	7.54	46,809,027	44,467,088	950	62,213,676	29,134,060	88,752	1,030
6	5	Madhya Pradesh	72,626,809	6	37,612,306	35,014,503	931	52,537,899	20,059,666	308,245	236
7	6	Tamil Nadu	72,147,030	5.96	36,137,975	36,009,055	996	37,189,229	34,949,729	130,058	555
8	7	Rajasthan	68,548,437	5.66	35,550,997	32,997,440	928	51,540,236	17,080,776	342,239	201
9	8	Karnataka	61,095,297	5.05	30,966,657	30,128,640	973	37,552,529	23,578,175	191,791	319
10	9	Gujarat	60,439,692	4.99	31,491,260	28,948,432	919	34,670,817	25,712,811	196,024	308
11	10	Andhra Pradesh	49,386,799	4.08	24,738,068	24,648,731	996	34,776,389	14,610,410	160,205	308
12	11	Odisha	41,974,218	3.47	21,212,136	20,762,082	979	34,951,234	6,996,124	155,707	269
13	12	Telangana	35,193,978	2.91	17,704,078	17,489,900	988	21,585,313	13,608,665	114,840	307
14	13	Kerala	33,406,061	2.76	16,027,412	17,378,649	1084	17,445,506	15,932,171	38,863	859
15	14	Jharkhand	32,988,134	2.72	16,930,315	16,057,819	948	25,036,946	7,929,292	79,714	414
16	15	Assam	31,205,576	2.58	15,939,443	15,266,133	958	26,780,526	4,388,756	78,438	397
17	16	Punjab	27,743,338	2.29	14,639,465	13,103,873	895	17,316,800	10,387,436	50,362	550
18	17	Chhattisgarh	25,545,198	2.11	12,832,895	12,712,303	991	19,603,658	5,936,538	135,191	189
19	18	Haryana	25,351,462	2.09	13,494,734	11,856,728	879	16,531,493	8,821,588	44,212	573
20	19	Delhi (UT)	16,787,941	1.39	8,887,326	7,800,615	868	944,727	12,905,780	1,484	11,297
21	20	Jammu and Kas	12,541,302	1.04	6,640,662	5,900,640	889	9,134,820	3,414,106	222,236	56
22	21	Uttarakhand	10,086,292	0.83	5,137,773	4,948,519	963	7,025,583	3,091,169	53,483	189
23	22	Himachal Prades	6,864,602	0.57	3,481,873	3,382,729	972	6,167,805	688,704	55,673	123
24	23	Tripura	3,673,917	0.3	1,874,376	1,799,541	960	2,710,051	960,981	10,486	350
25	24	Meghalaya	2,966,889	0.25	1,491,832	1,475,057	989	2,368,971	595,036	22,429	132
26	25	Manipur	2,855,794	0.24	1,438,687	1,417,107	985	1,899,624	822,132	22,327	128
27	26	Nagaland	1,978,502	0.16	1,024,649	953,853	931	1,406,861	573,741	16,579	119
28	27	Goa	1,458,545	0.12	739,140	719,405	973	551,414	906,309	3,702	394
29	28	Arunachal Prade	1,383,727	0.11	713,912	669,815	938	1,069,165	313,446	83,743	17
30	29	Puducherry (UT)	1,247,953	0.1	612,511	635,442	1037	394,341	850,123	479	2,598
31	30	Mizoram	1,097,206	0.09	555,339	541,867	976	529,037	561,997	21,081	52
32	31	Chandigarh (UT)	1,055,450	0.09	580,663	474,787	818	29,004	1,025,682	114	9,252
33	32	Sikkim	610,577	0.05	323,070	287,507	890	455,962	151,726	7,096	86
34	33	Andaman and Ni	380,581	0.03	202,871	177,710	876	244,411	135,533	8,249	46
35	34	Dadra and Naga	343,709	0.03	193,760	149,949	774	183,024	159,829	491	698
36	35	Daman and Diu (	243,247	0.02	150,301	92,946	618	60,331	182,580	112	2,169
37	36	Lakshadweep (U	64,473	0.01	33,123	31,350	946	14,121	50,308	32	2,013
38	-	Total (India)	1,210,854,977	100	623,724,248	586,469,174	943	833,087,662	377,105,760	3,287,240	382

Now that I had received my data, my new goal was to transfer it to a data frame. To do that, I had to download this data as a .csv file.

Next step was to open it in RStudio and work my magic.

First, I had to set the working directory into the place where I had saved the .csv file:

```
Set the working directory
##{r}
setwd("D:/Mahitha/DataScience/Hw7/")
```

Next, I had to convert this .csv file into a data frame:

```
Convert the downloaded csv file into a dataframe
##{r}
loadDF <- function() {
  if(!exists("demographics")) {
    demographics <- read.csv("Data Science - Demographics of India - Sheet1.csv", header = TRUE)
  }
}
```

And just to prove that demographics is indeed a data frame:

```
To show that this is in fact a dataframe, we can call its class type:
##{r}
class(demographics)
```

```
[1] "data.frame"
```

Also, I can print out its values and show that all of its data is present from the .csv file:

```
To show that this is a dataframe, I can also print it out and show it to you:
##{r}
print(demographics)
```

Rank <fctr>	State.UT <fctr>	Population.48. <dbl>	Percent.... <dbl>	Male <fctr>	Female <fctr>
1	Uttar Pradesh	199,812,341	16.50	104,480,510	95,331,831
2	Maharashtra	112,374,333	9.28	58,243,056	54,131,277
3	Bihar	104,099,452	8.60	54,278,157	49,821,295
4	West Bengal	91,276,115	7.54	46,809,027	44,467,088
5	Madhya Pradesh	72,626,809	6.00	37,612,306	35,014,503
6	Tamil Nadu	72,147,030	5.96	36,137,975	36,009,055
7	Rajasthan	68,548,437	5.66	35,550,997	32,997,440
8	Karnataka	61,095,297	5.05	30,966,657	30,128,640
9	Gujarat	60,439,692	4.99	31,491,260	28,948,432
10	Andhra Pradesh	49,386,799	4.08	24,738,068	24,648,731

1-10 of 37 rows | 1-6 of 11 columns

Previous 1 2 3 4 Next

There are 4 pages of data, and going through each of them will prove that I have successfully created a data frame from a URL.

As you can see, this data looks very clean, and the only thing I would personally do to clean this data is to remove the "Rank" row which is at the far left. Although it may have some symbolic meaning - the order of states by population from greatest to least - it is not necessary to have a separate row to show that. It can quite simply be written as a remark and also can be very easily seen by someone analyzing this data.