Mahitha Valluru

Homework 7

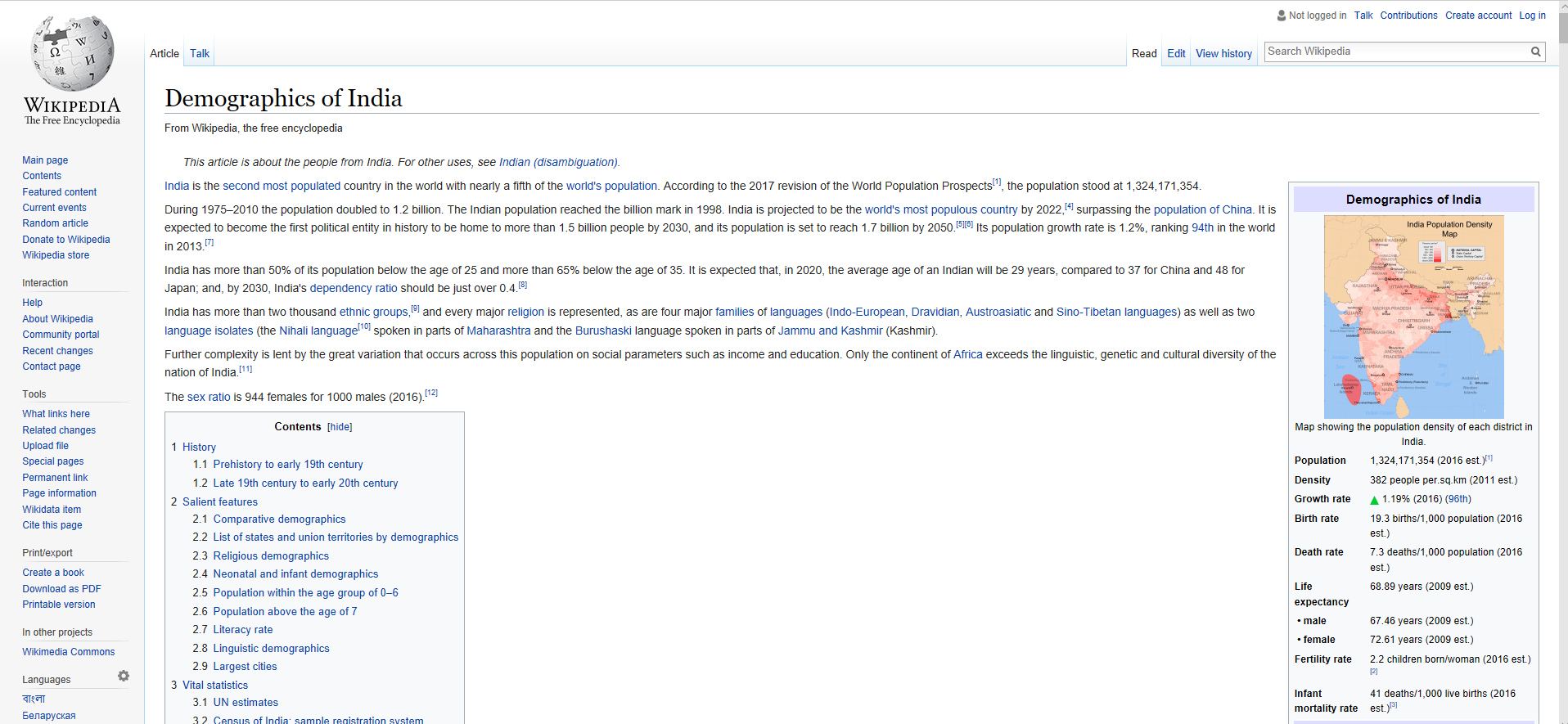
DS4100: Data Science

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

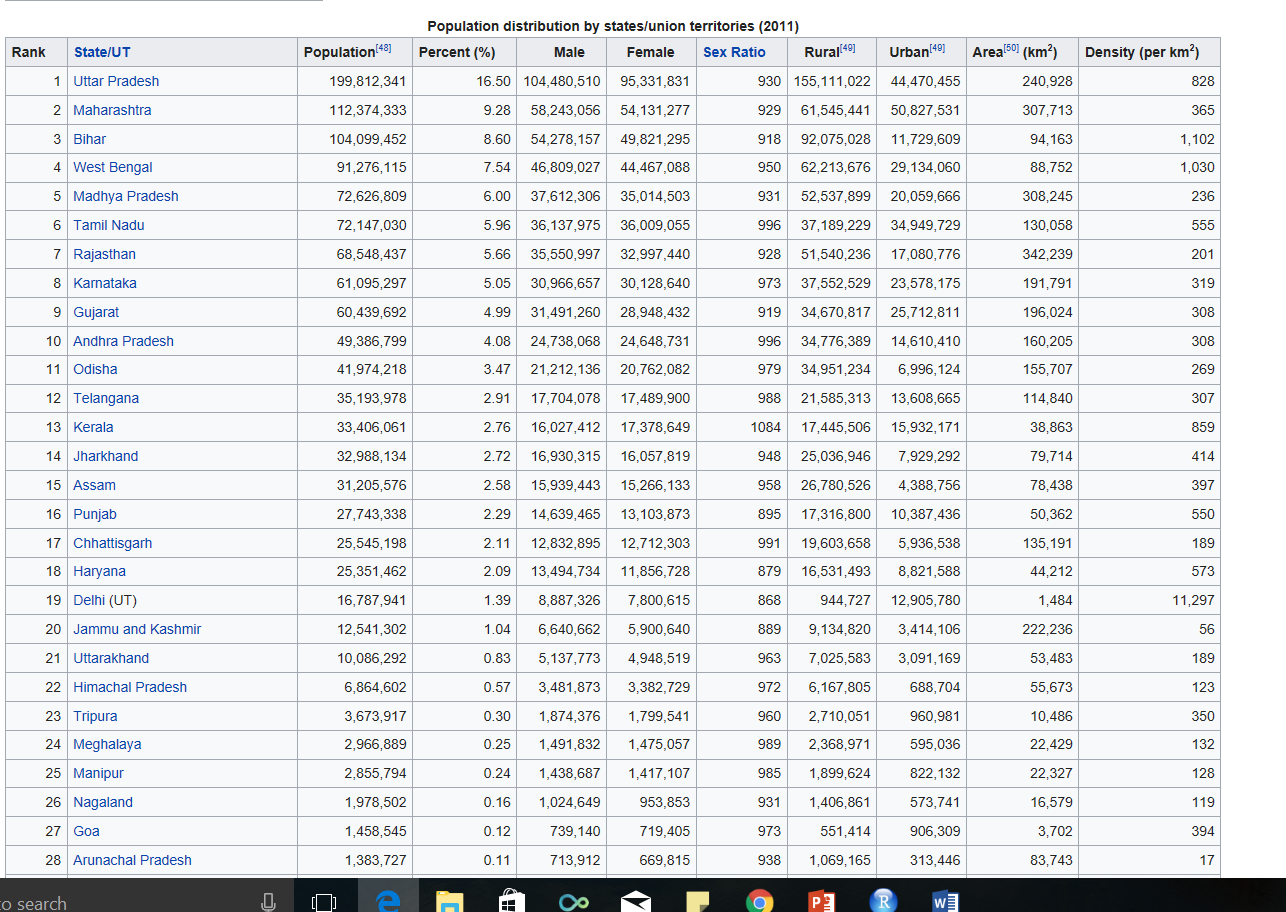
Assignment 7 indicates that I must choose a web scraping toolkit to scrape the data from a website of my choice.

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:



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



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

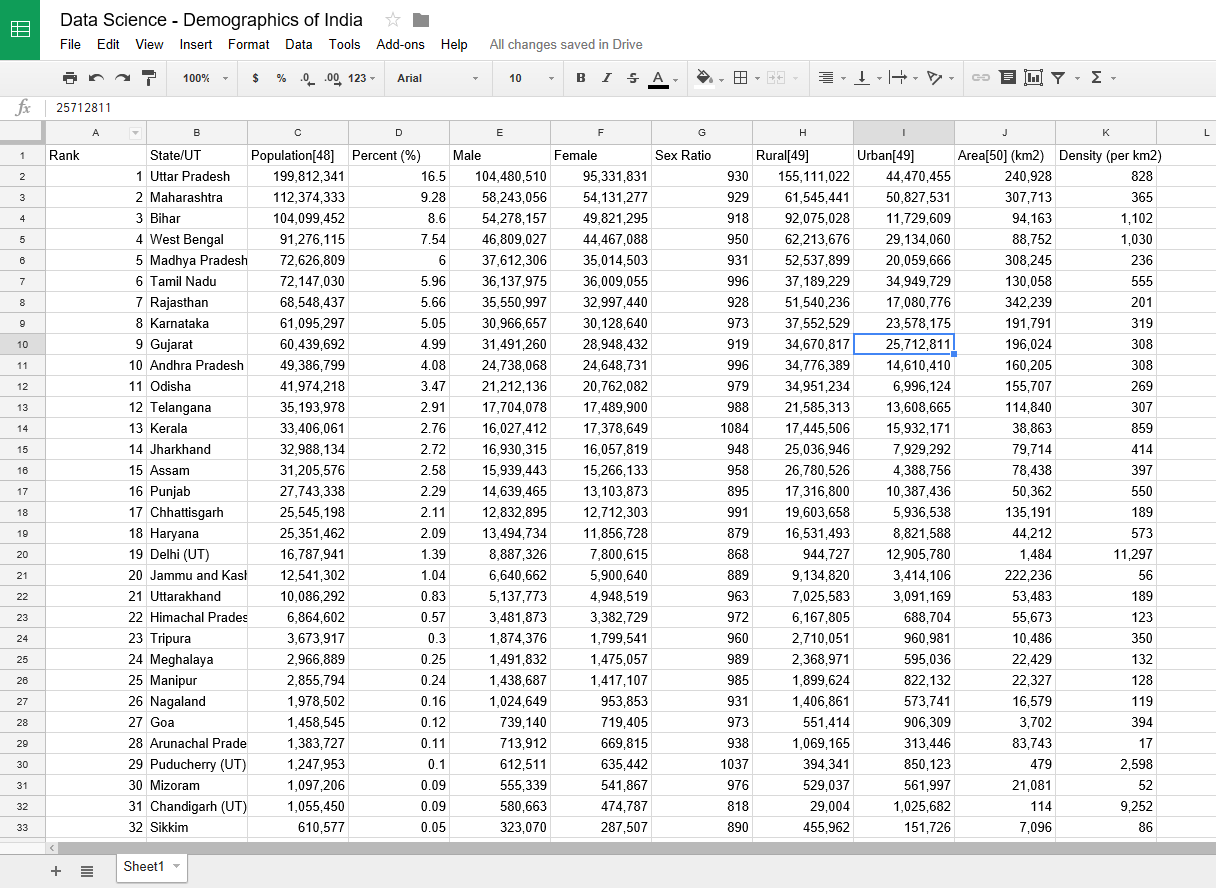
When using the Google Sheets web scraping toolkit, I used the ImportHtml() function.

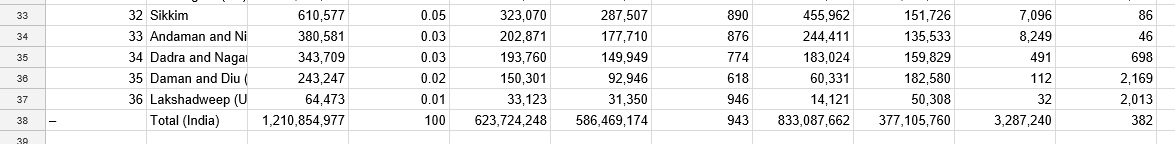
To use it, I needed to enter some information into the function in the top left-hand corner of the cells, so in cell A1:

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:

And received this as a result:

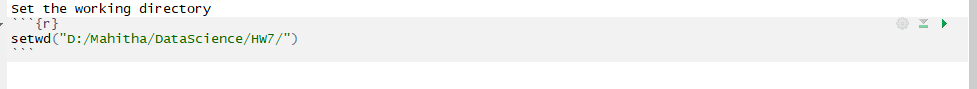




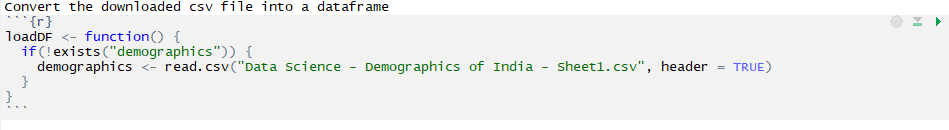
Now that I had received my data, my new goal was to transfer it to a data frame. To do that, I had to download this data as a .csv file.

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

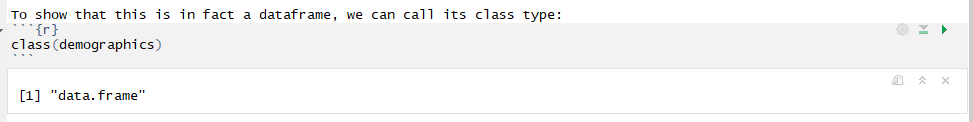
First, I had to set the working directory into the place where I had saved the .csv file:



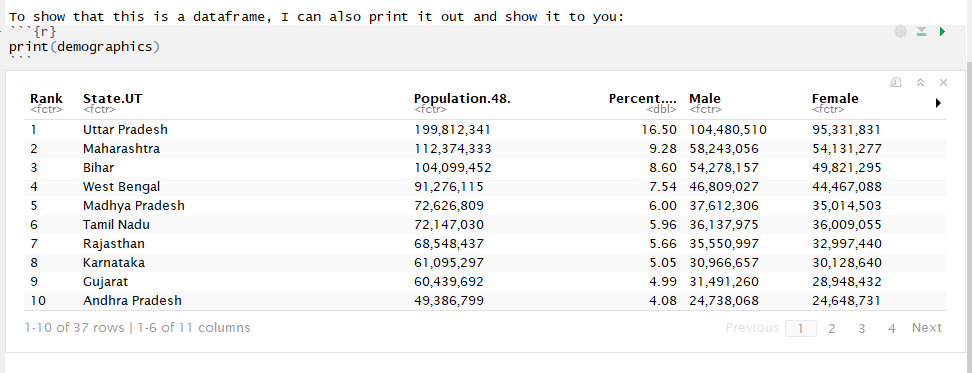
Next, I had to convert this .csv file into a data frame:



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



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

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.