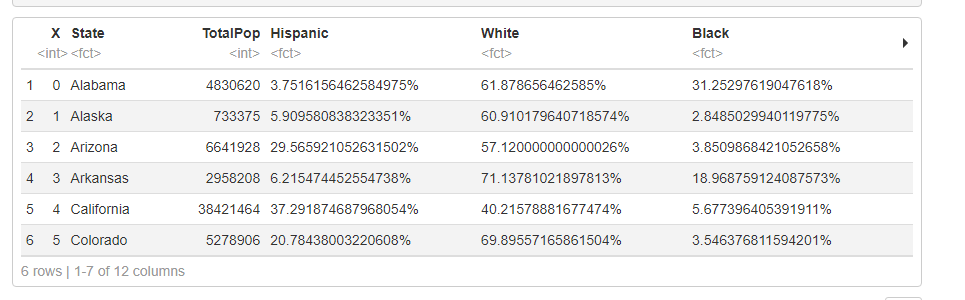
**The results of the project**

We got the dataset from the Census Bureau, which collects census data and finds interesting insights from it.

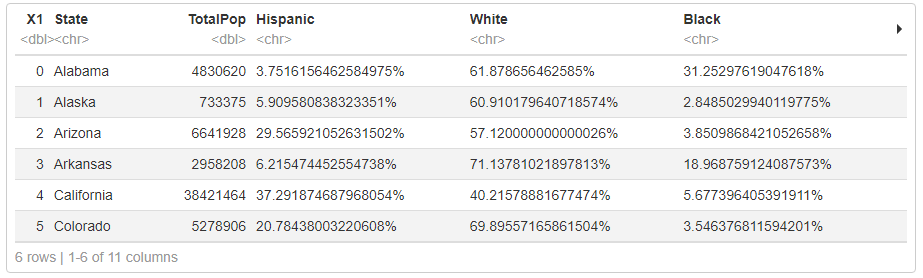
So at first, before we make any data analysis, we need to prepare our dataset into reasonable shape and form so that we can perform our analysis.

+ First, we need to inspect some of the csv files to know what variables and observations we need to attention.

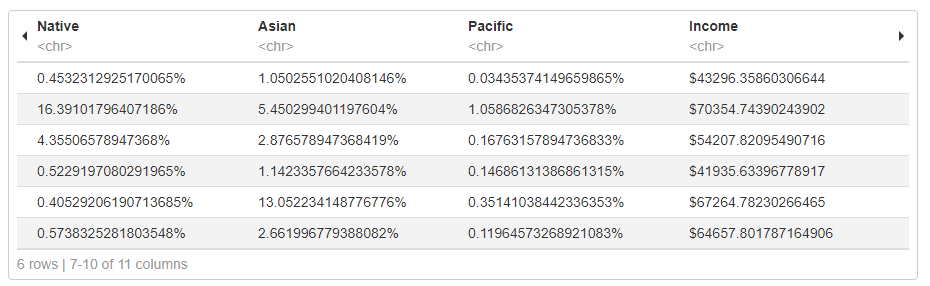


Because we have numerous dataset that follows this structure: “states\_.csv” so we will create a list of them and combine their rows together.

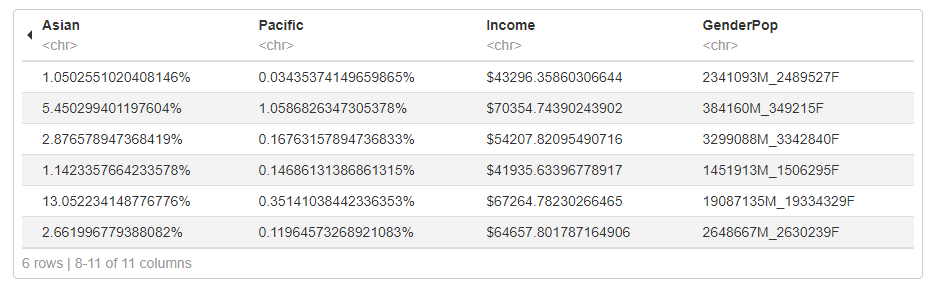
Our dataset is so wide. So I split it into 3 parts.



Part 1



Part 2



Part 3

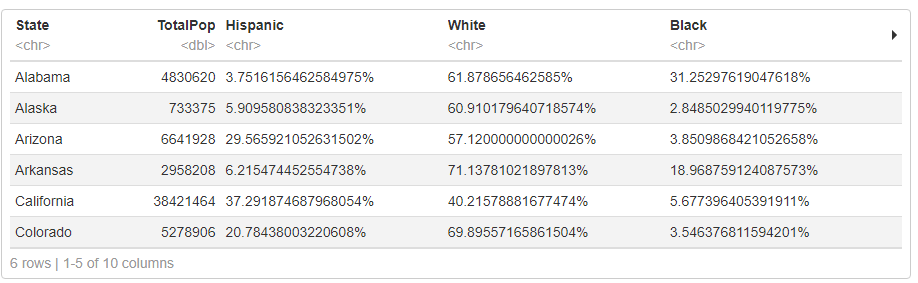
Also, it is better to check the datatypes of columns to find out which columns we need to fix.



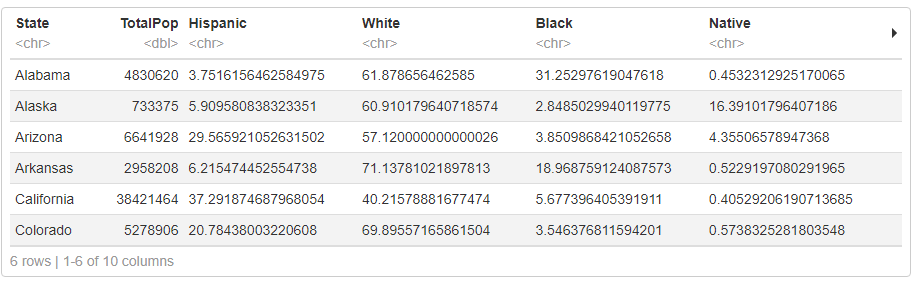
As we can see, this dataset contains myriad issues below:

* Hispanic, White, Black, Native, Asian, Pacific, Income and GenderPop must be transferred to numeric datatype. The method is to make sure all of these columns to contain only number and then convert it to numeric at once.
* GenderPop must be split into columns for males and columns for female.
* X1 column is useless. We must omit it.
* Check for duplicates.

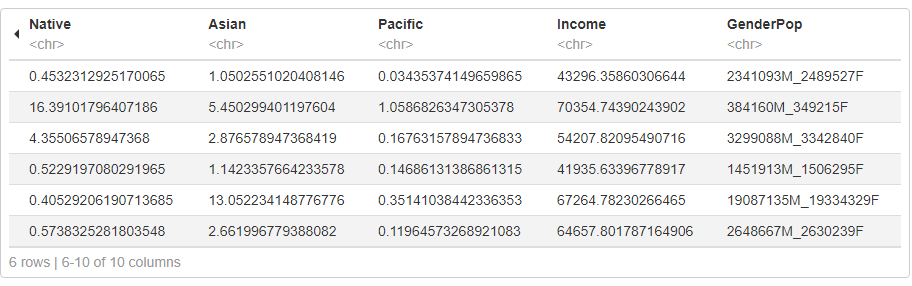
+ Second, begin with something easy, we will omit the X1 column.



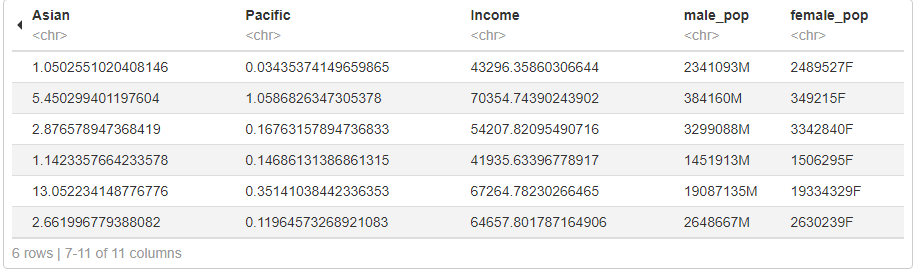
+ Third, since the races columns (Hispanic, White, Black,…) have the percentage signs in their values. In order to convert it into numeric, we must omit all the special character and keep the numbers only.



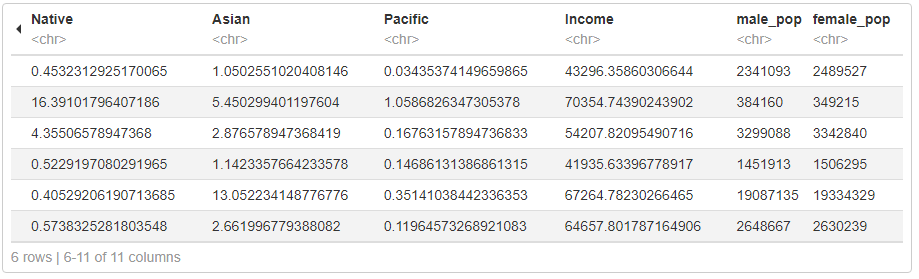
+ Also, the Income column also contains the $ sign. So we can omit too.



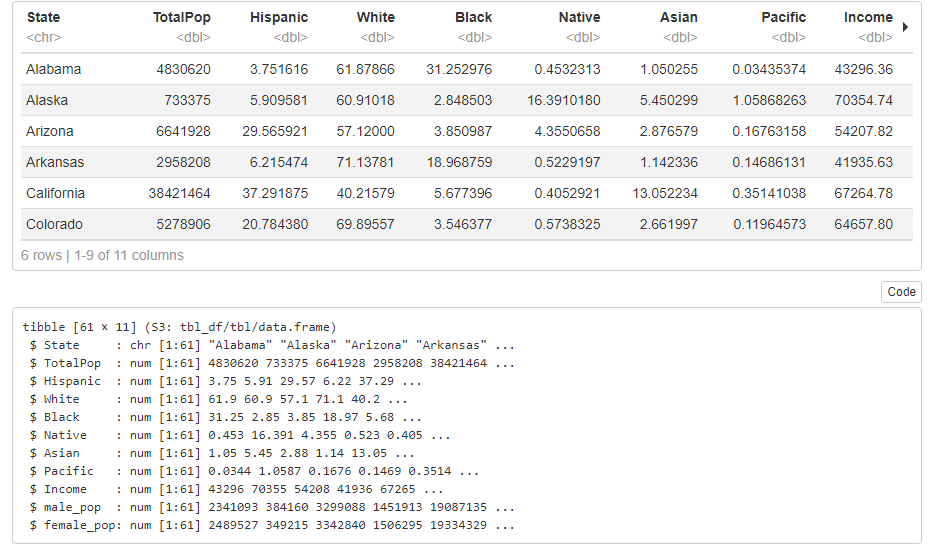
+ Fourth, we will begin to separate the GenderPop column into the male column and female column.



Then we will omit the M and F character in both columns.

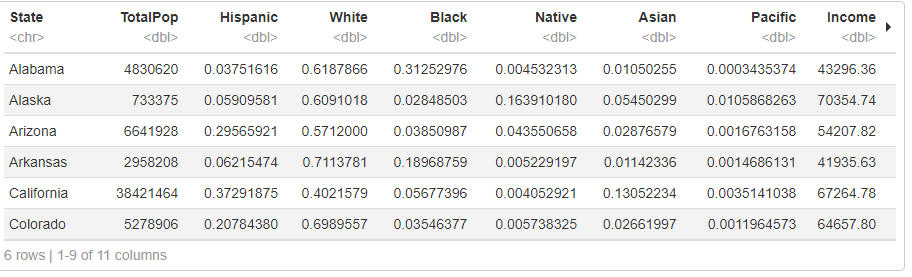


+ Finally, we can convert these columns into numeric datatype.



This is the result.

+ Because the races columns demonstrate the percentage of each race in each state. Our columns need to transpose into decimal values.



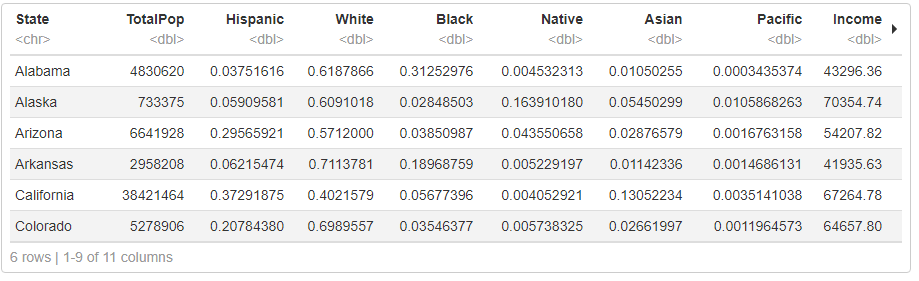
+ Now we begin to check for duplicates in our dataset.



As we can see, there are 9 cases that duplicates. So we will delete them.



This is the result after we delete duplicates and double check again. There aren’t any duplicates.



Finally, this is our result. This data are ready for analysis.