# Prep Exercise (PE03) Data Cleansing and Munging

### General Instructions

1. For this exercise you will answer all of the questions in this document and turn it in to Blackboard.
2. Review from last week
   1. Make a data frame data.frame( )
   2. Row index of max/min which.max( ) which.min( )
   3. Sort value or order rows sort( ) order( )
   4. Descriptive statistics mean( ) sum( ) max( )
3. Getting Started:
   1. Often, in data science, when you get a dataset, it is not in the format you want. So, you have to use R code to refine the dataset into something more useful. As Chapter 6 of Introduction to Data Science mentions, this is called “data munging.” In this prep exercise, you will read in a dataset from the web and work on that dataset (in a dataframe) so that it can be useful. Once the dataset is prepped (and checked for validity) you will use the dataframe in your homework to explore the distribution of a variable within the dataset.
   2. IFF (if and only if) the web is not available to read we have a copy of the dataset in BB for you to use.

# IST 687, Standard Homework Heading

#

# Student name: Thadhani Hitesh Chandrakumar

# Homework number: PE03

# Date due: Wed 11th Sep 2019 11:59PM

#

# Attribution statement: (choose the statements that are true)

# 1. I did this work by myself, with help from the book and the professor

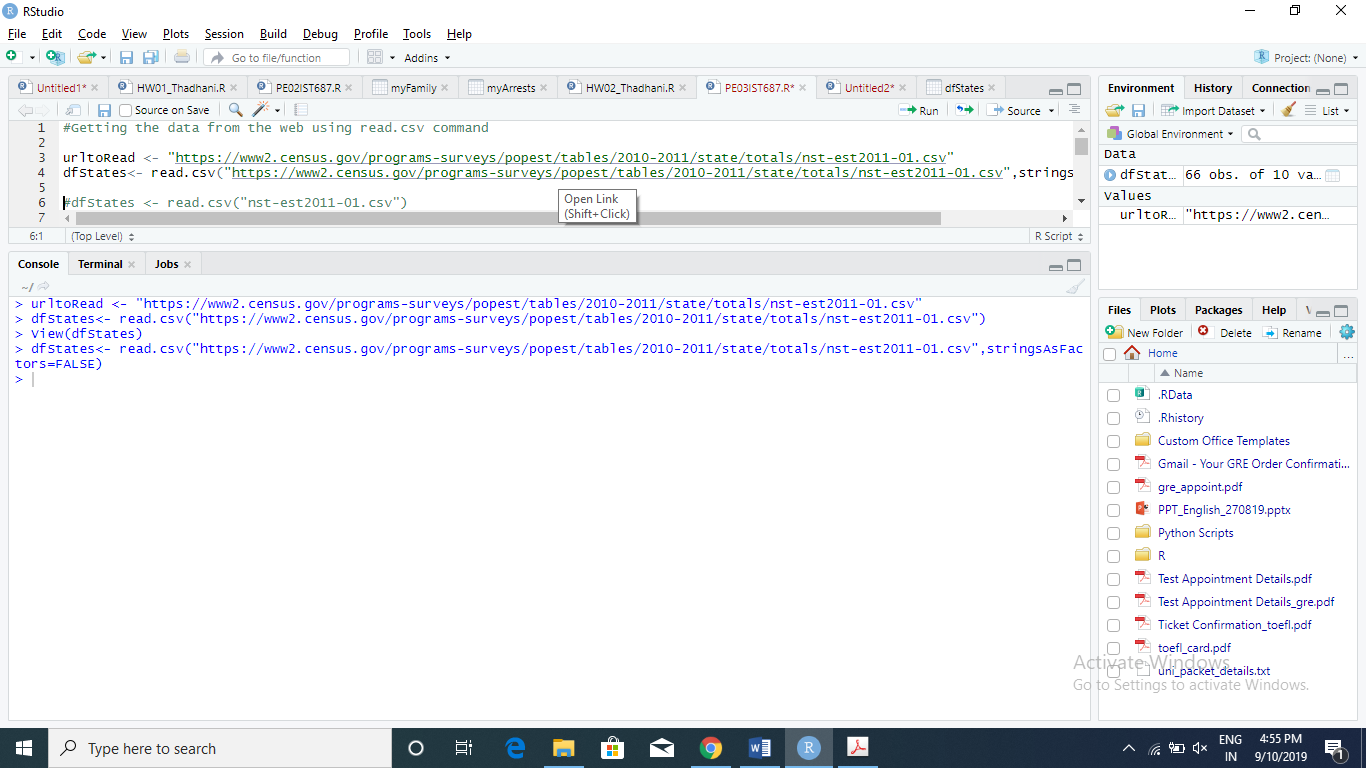
# 2. I did this homework with help from the book and the professor and these Internet sources: <provide the urls>

# 3. I did this homework with coaching from <Name of another student> but did not cut and paste any code

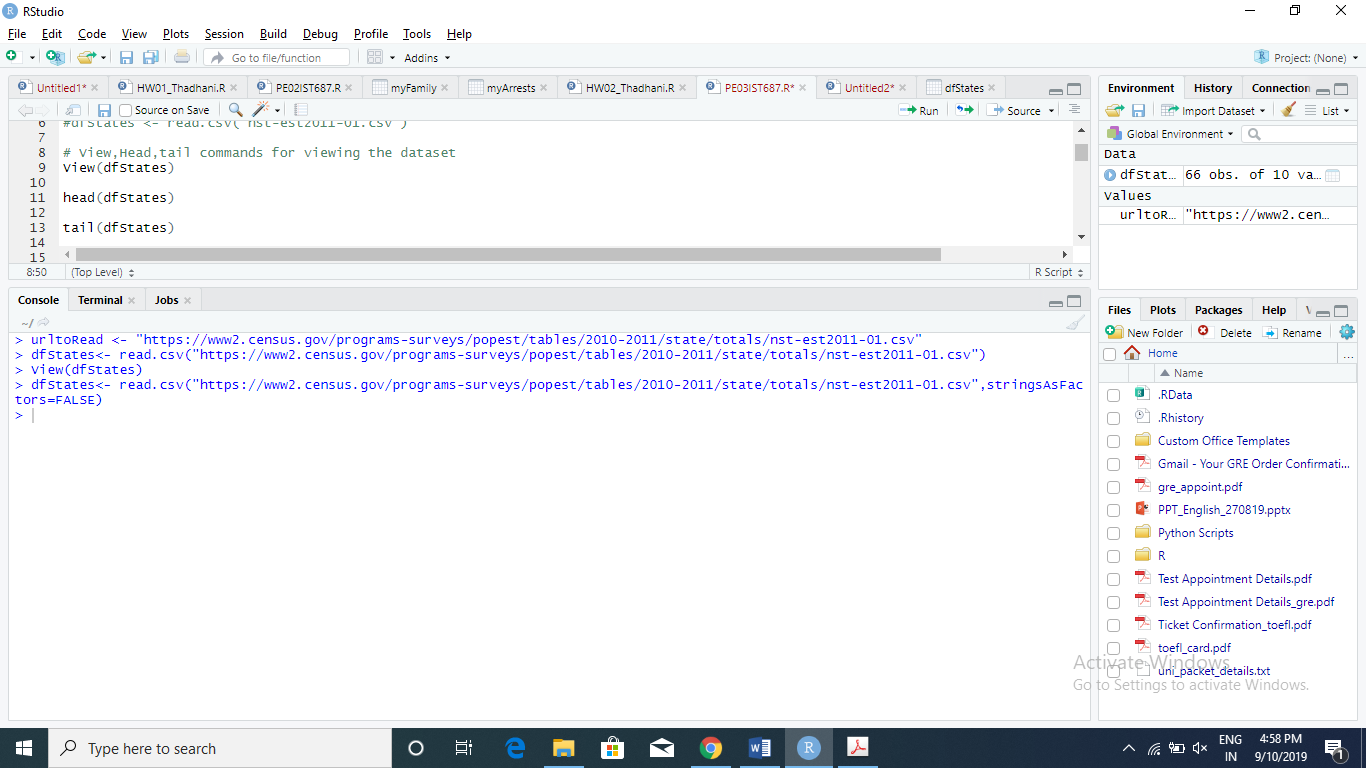
### Prep Exercise

1. **Getting the Data**
2. Use R code to read directly from a URL on the web. *Hint: use read.csv and url() to read the file from the web*.
3. Store the dataset into a new dataframe, called dfStates. Hint: Use stringsAsFactors=FALSE.
4. The URL is:

<https://www2.census.gov/programs-surveys/popest/tables/2010-2011/state/totals/nst-est2011-01.csv>

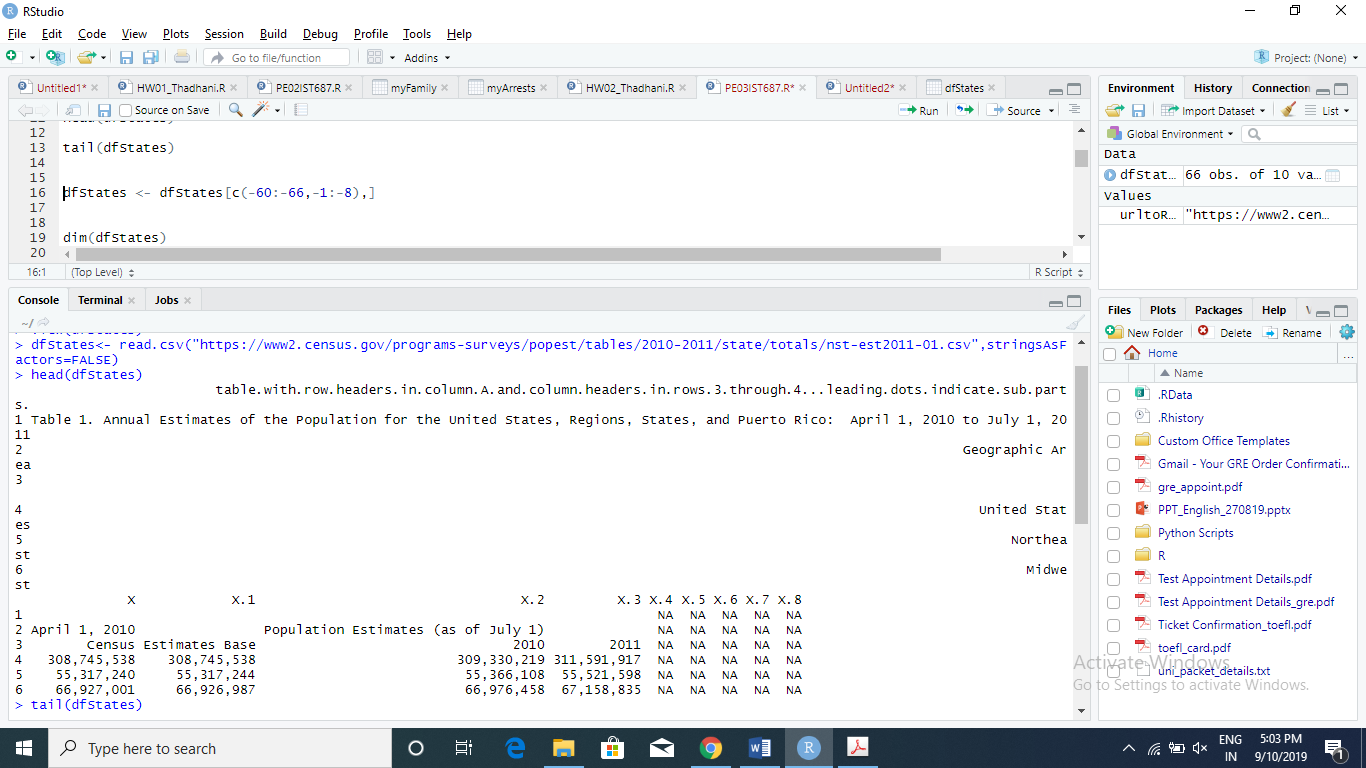


1. **Clean Up the Dataframe**
2. Use View( ), head( ), and tail( ) to examine the data frame. Briefly describe what each of the commands show you about a dataframe in general and as it relates to this dataset.

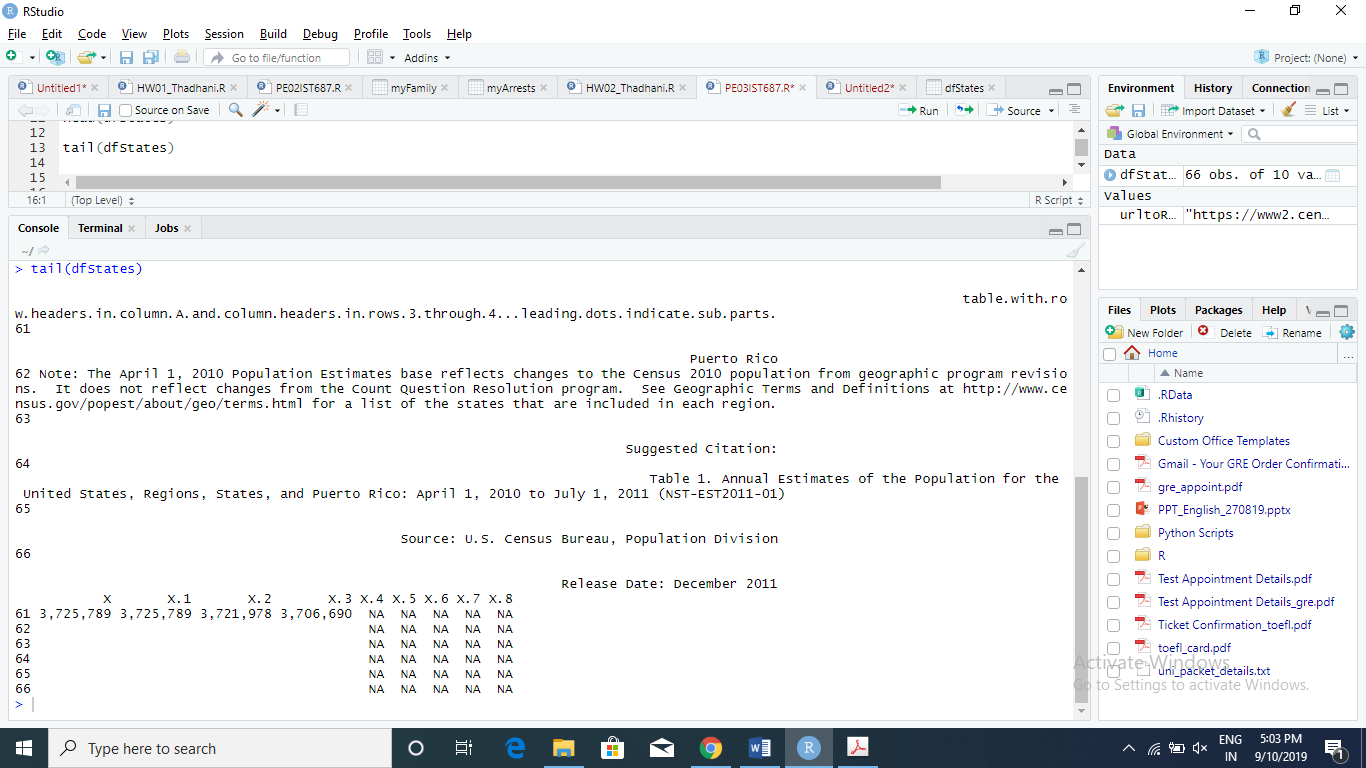


View() : Tabular format view of the dataset rows and columns

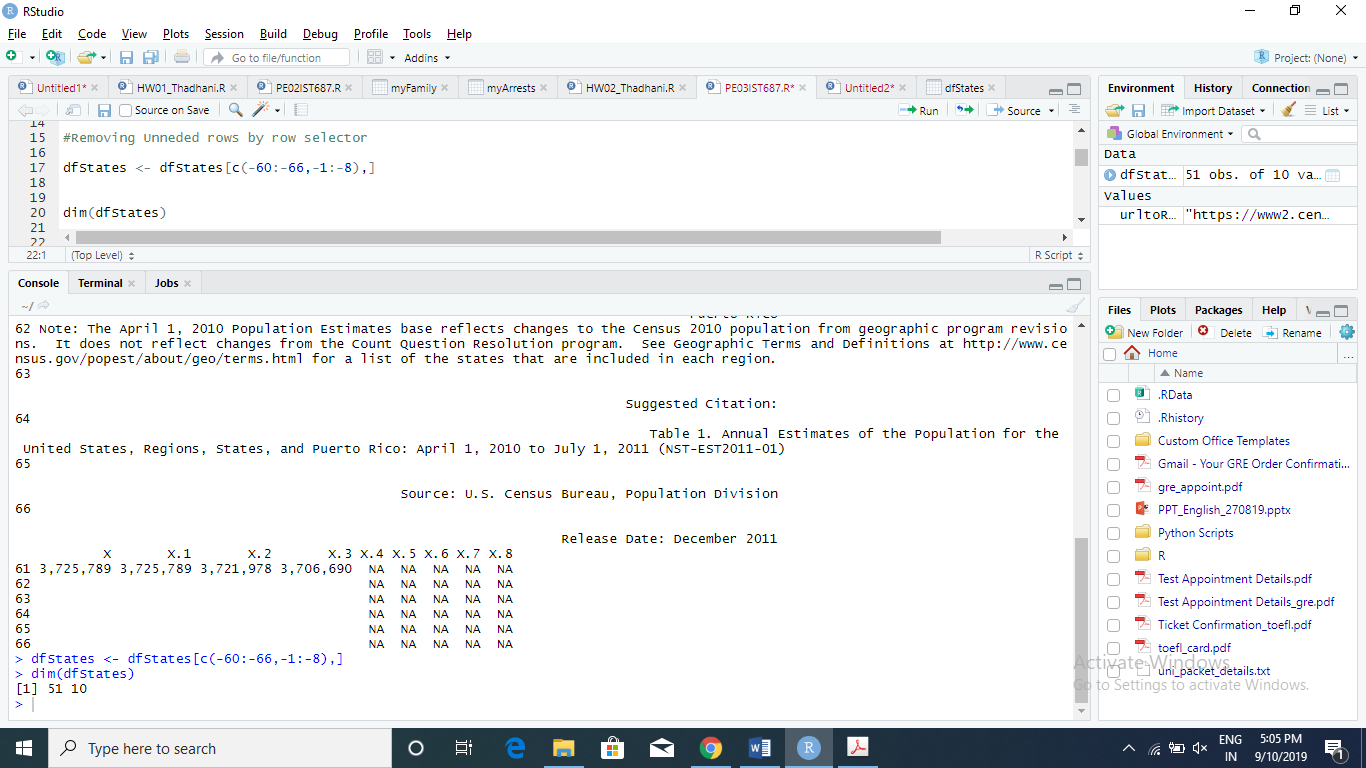
Head() : Gives the 1st 6 rows of the dataset



Tail() : Gives the last 6 rows of the dataset



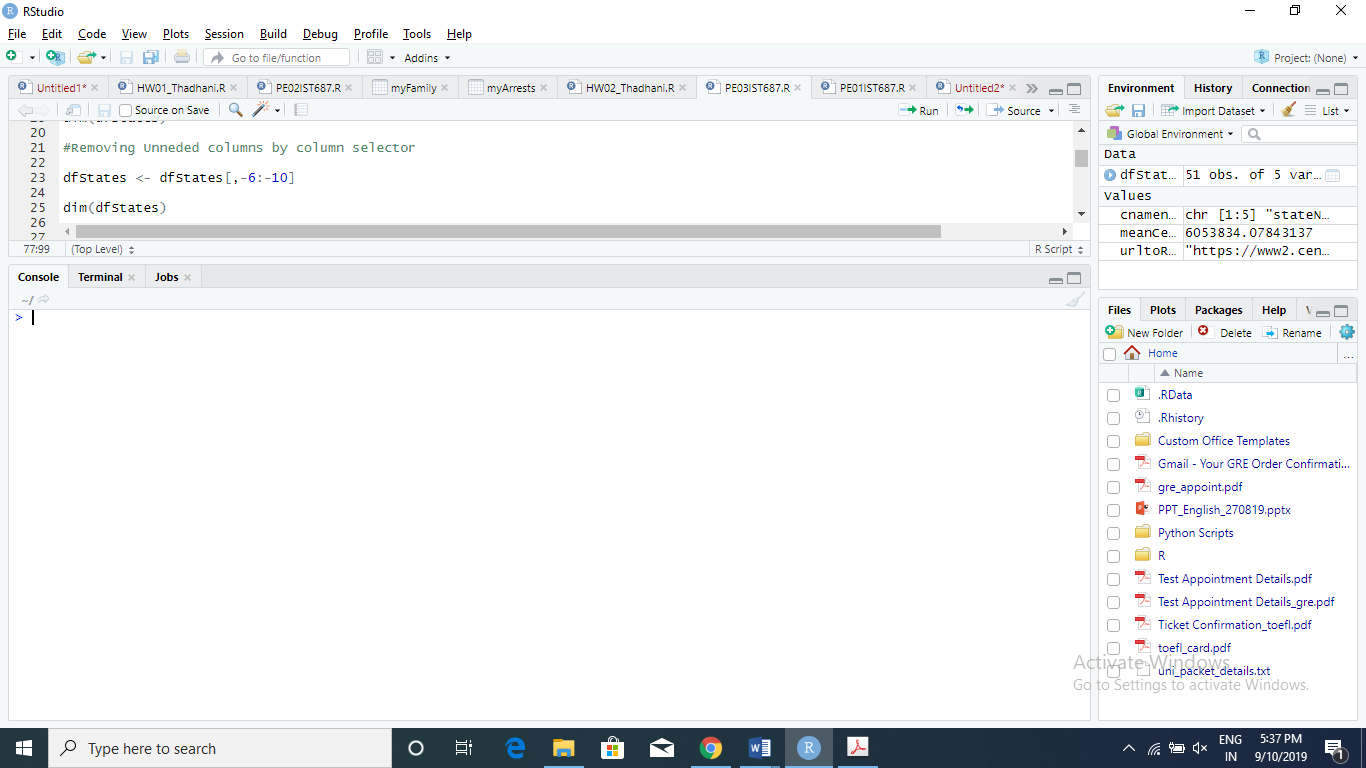
1. Remove unneeded rows by using the minus sign in the row selector of the [ , ] subsetting method. *Hint: Start by removing extra rows that appear* ***at the end*** *of the data set.*



1. Use the dim() command to make sure there are exactly 51 rows (one per state + the district of Columbia).

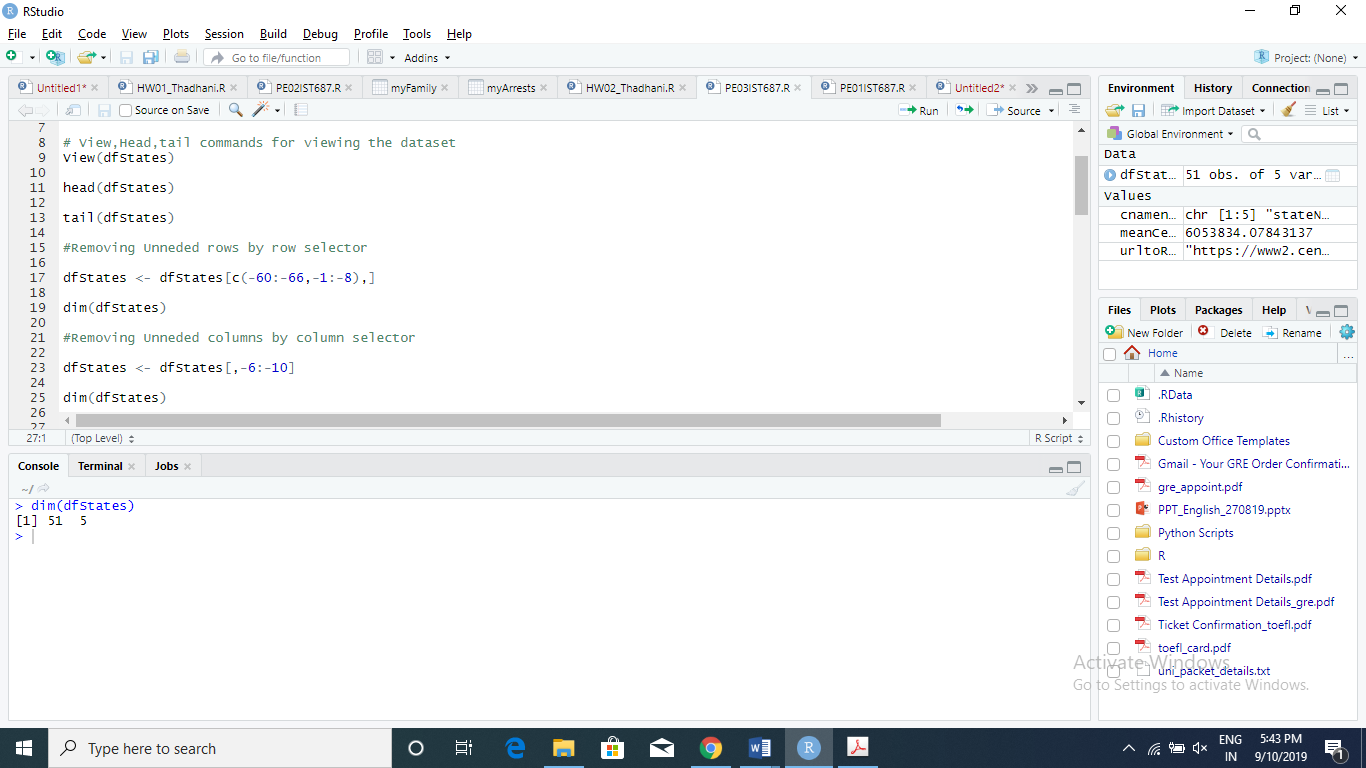
The output above shows 51 rows exactly with 10 variables.

1. Remove unneeded columns by using the minus sign in the column selector of the [ , ] subsetting method.

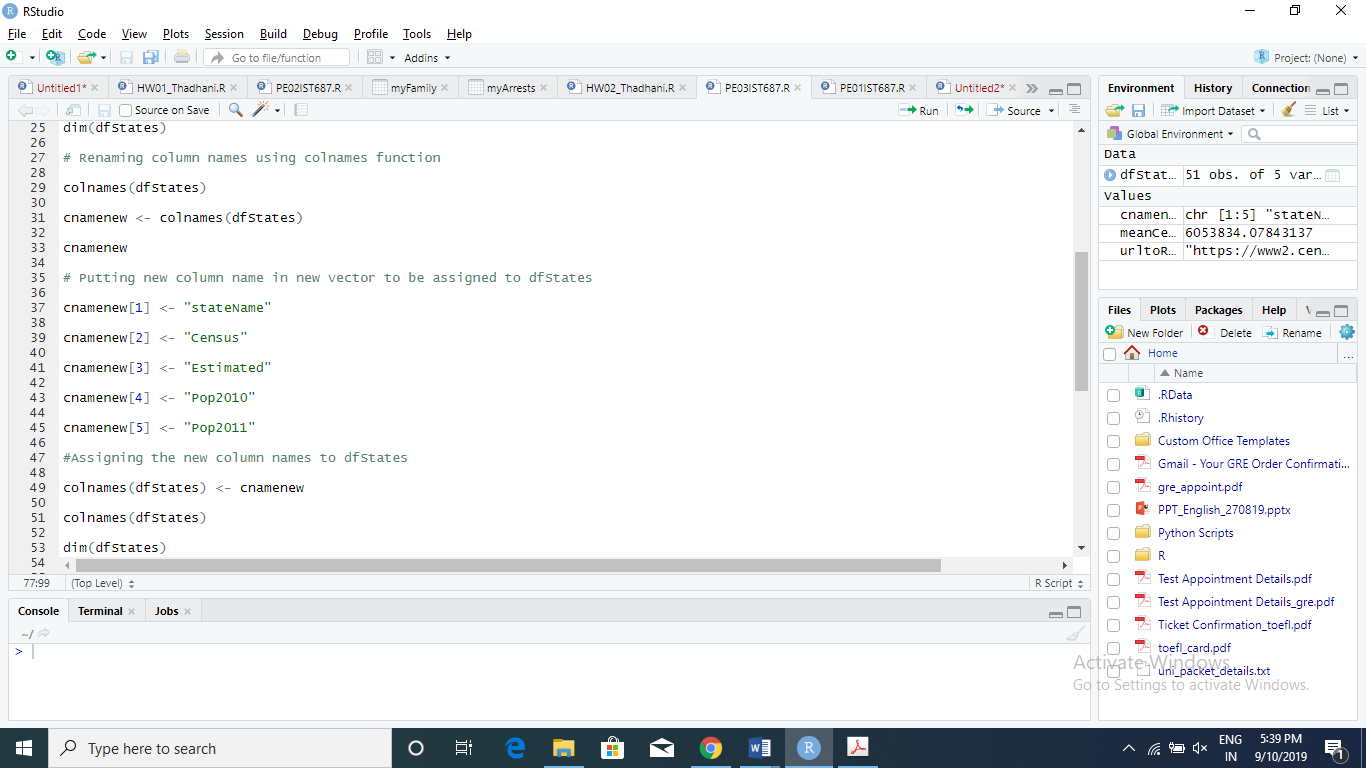


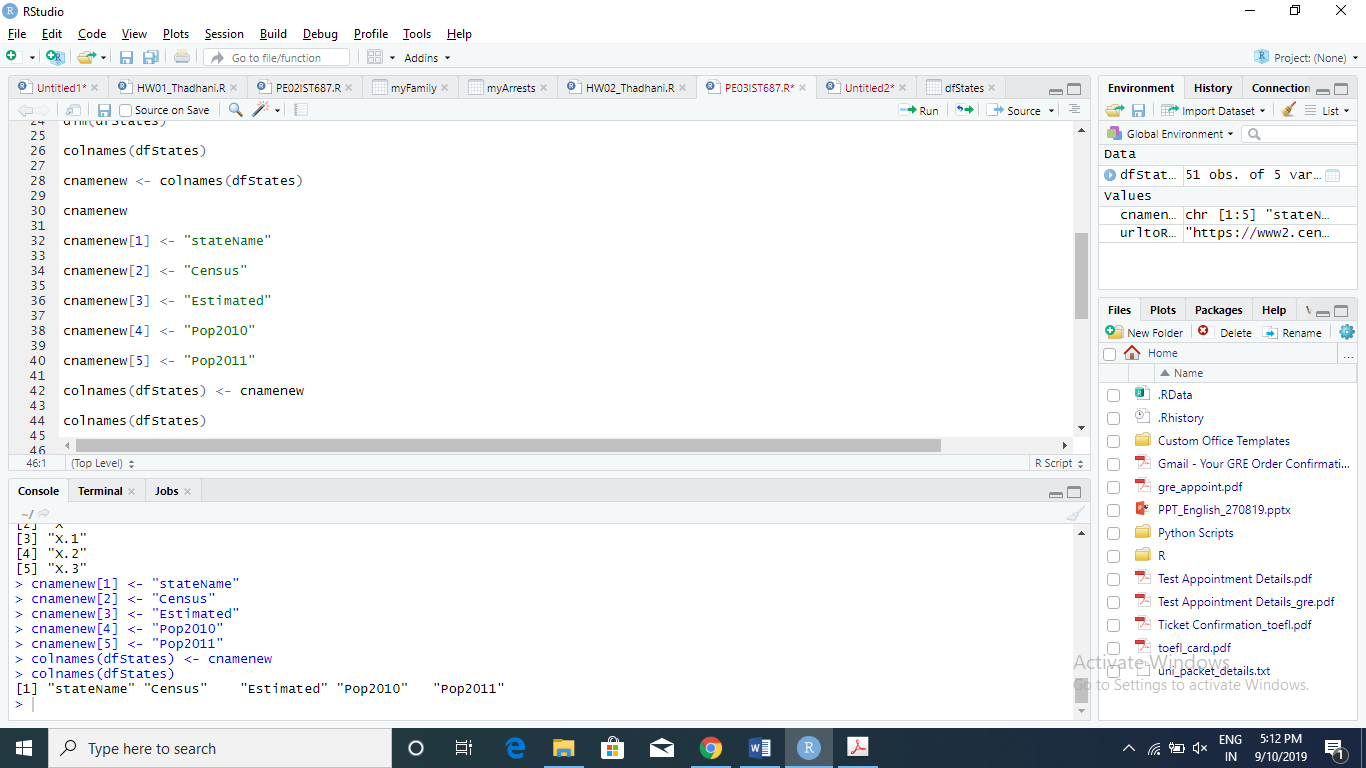
1. Use the dim() command to make sure there are precisely five columns.

**The Screenshot below shows that there are exactly 51 rows and 5 columns using dim() command.**

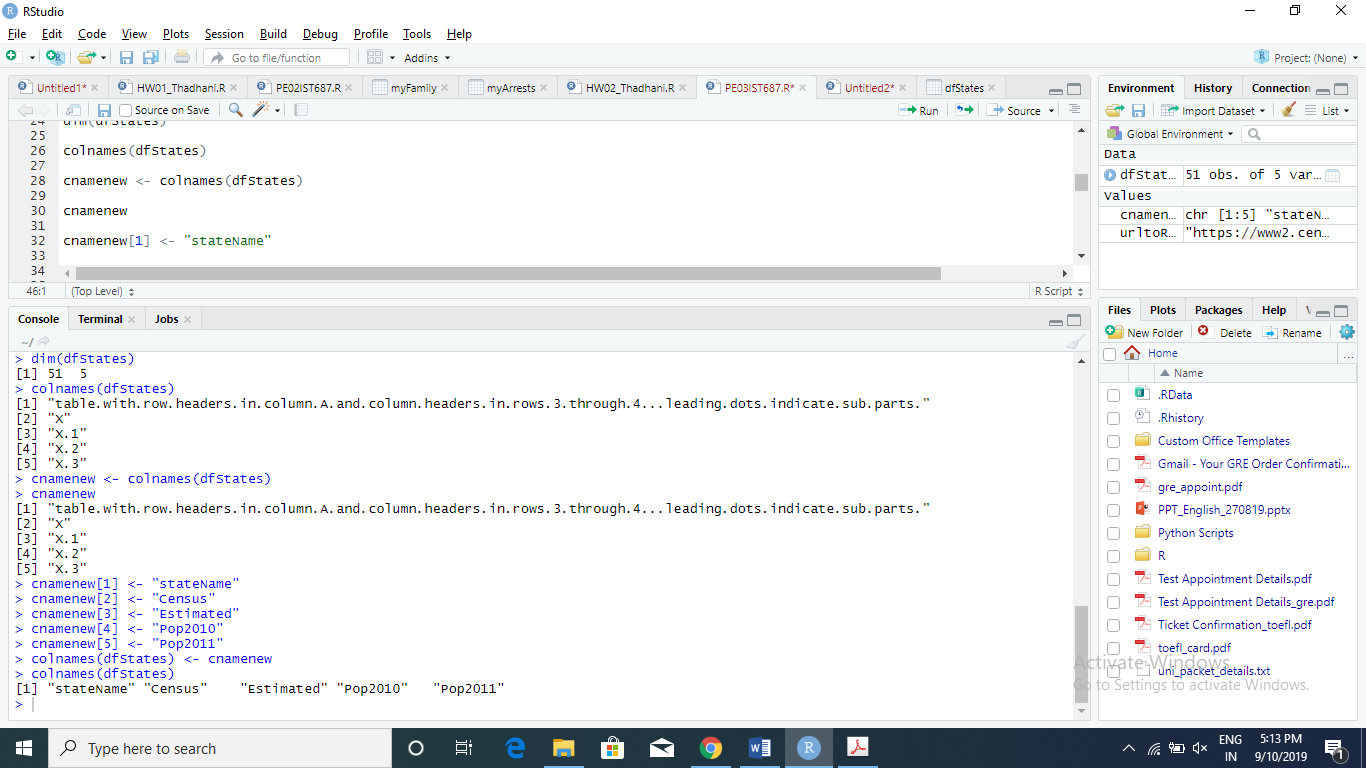


1. **Let’s add some meaningful metadata!**
2. Rename the columns with the following names: stateName, Census, Estimated, Pop2010, Pop2011. *Hint: use colnames( )*





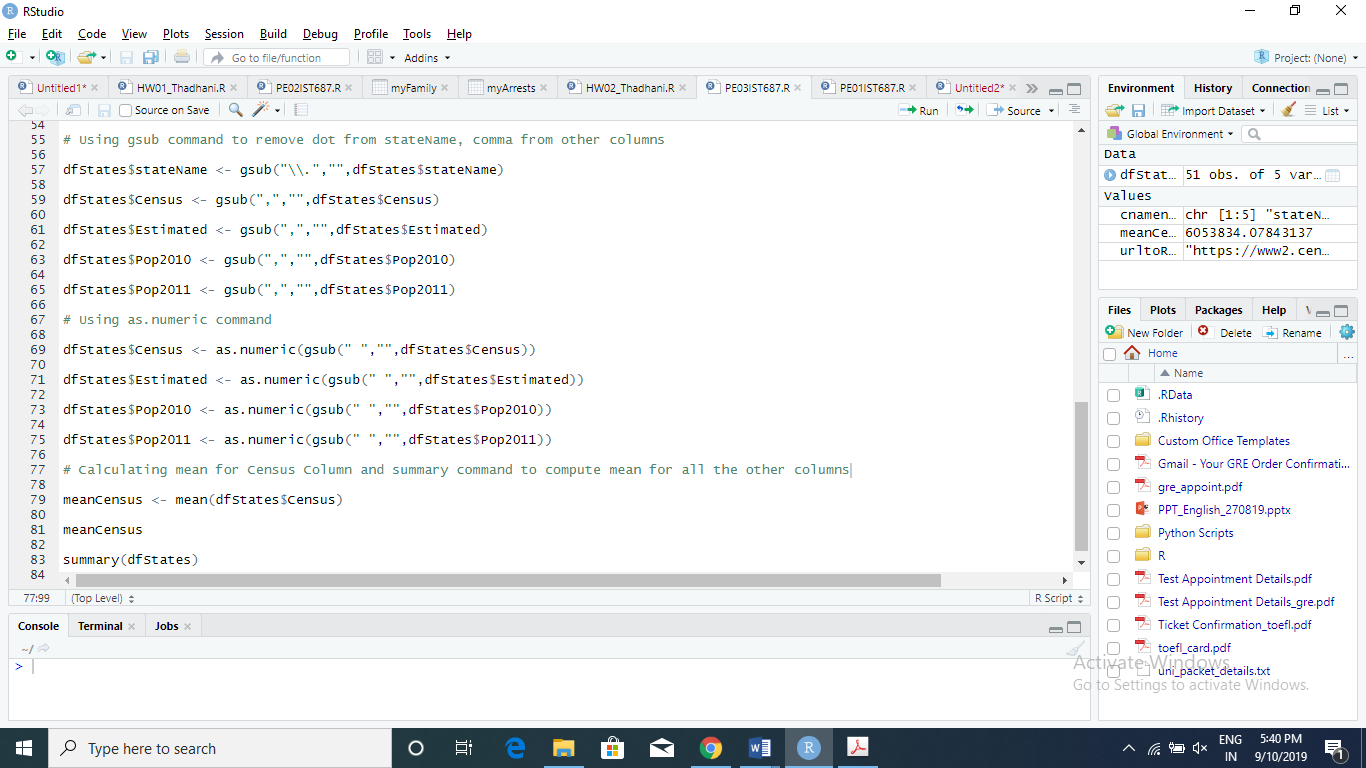
Code for the change of column names



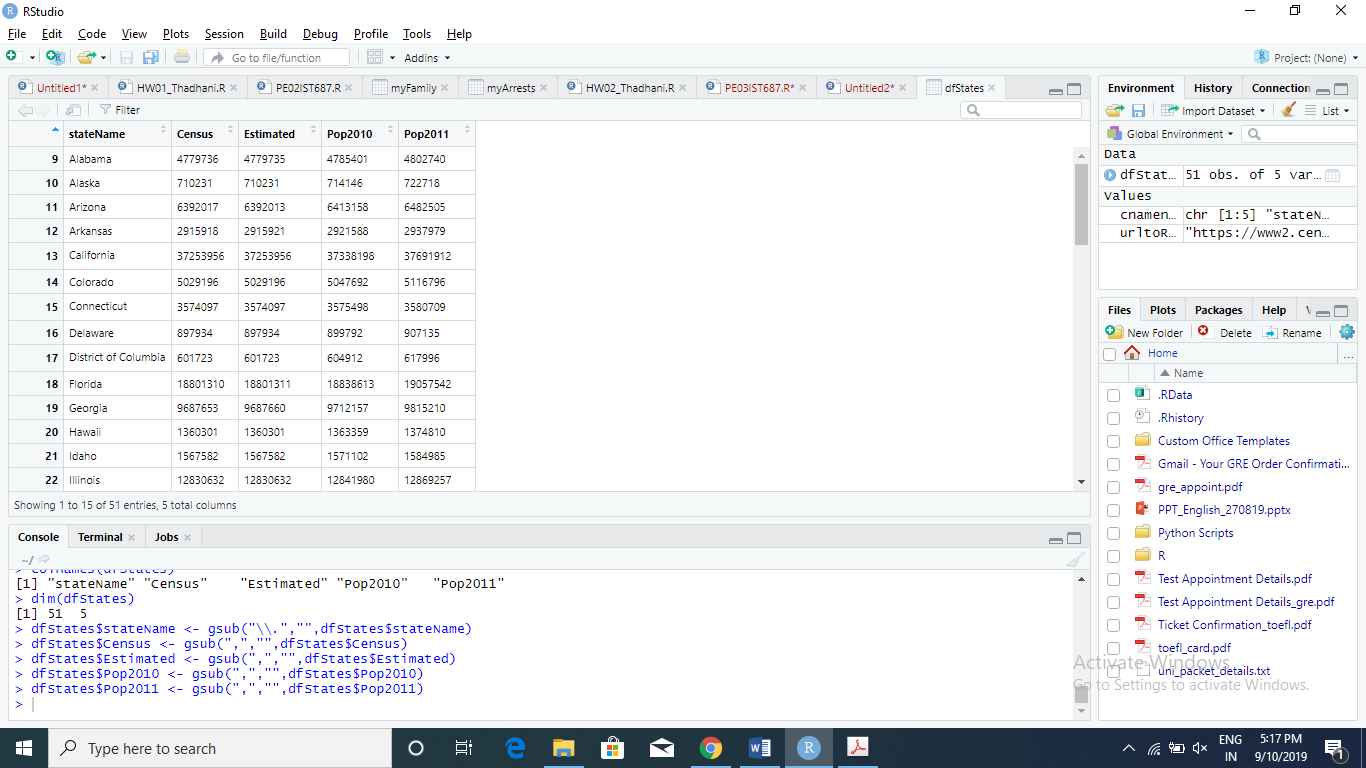
Output of colnames containing changed colnames

1. **More cleansing!** 
   1. Use gsub(",", "", vectorName) to remove the commas from each column of numeric data. Place the converted results back into the data frame.
   2. Use as.numeric( ) to coerce each numeric data column into numbers. Place the converted results back into the data frame
   3. Calculate the mean of the 4 numeric variables and fill in the table below:

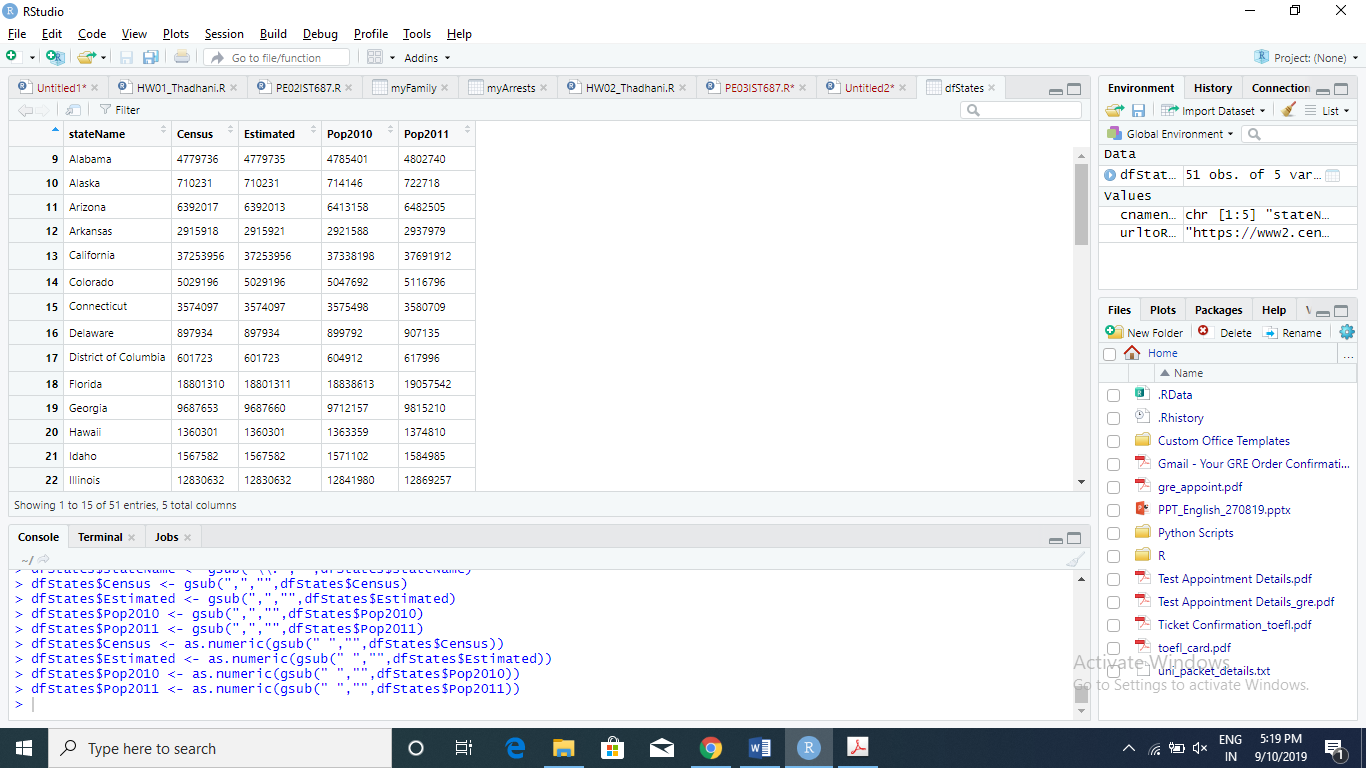
|  |  |
| --- | --- |
| **Census** | 6053834 |
| **Estimated** | 6053834 |
| **Pop2010** | 6065298 |
| **Pop2011** | 6109645 |

**Code for gsub to remove commas, as.numeric and mean of the columns** 

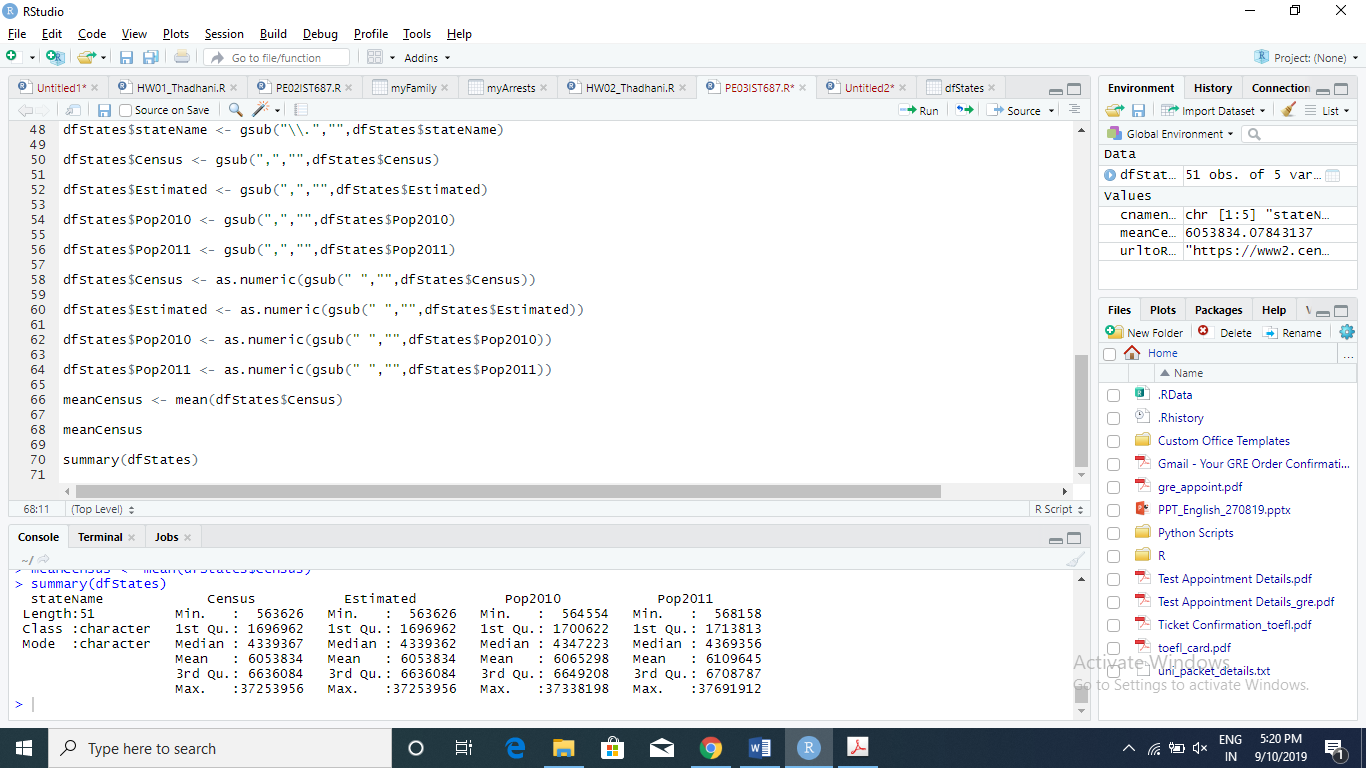
**View of dfStates after gsub command removing the commas from the columns**



**Output after as.numeric command execution for the columns**



**Mean finding using mean command and for the other columns summary(dfStates) can give the mean of all the columns which are populated in the table**



1. **List any additional resources you used here.**

Used Book Introduction to Data Science by Saltz & Stanton Chap 6 Data Munging

1. **Be sure to save your R file as this will become the starting code for your homework.**

***You must submit all Prep Exercises to blackboard prior to the deadline specified for each assignment.*** PE assignments are due on the evening prior to the lecture class. Late PE assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*