# Prep Exercise (PE07) Data Prep for Visualizations using Map Data

### General Instructions

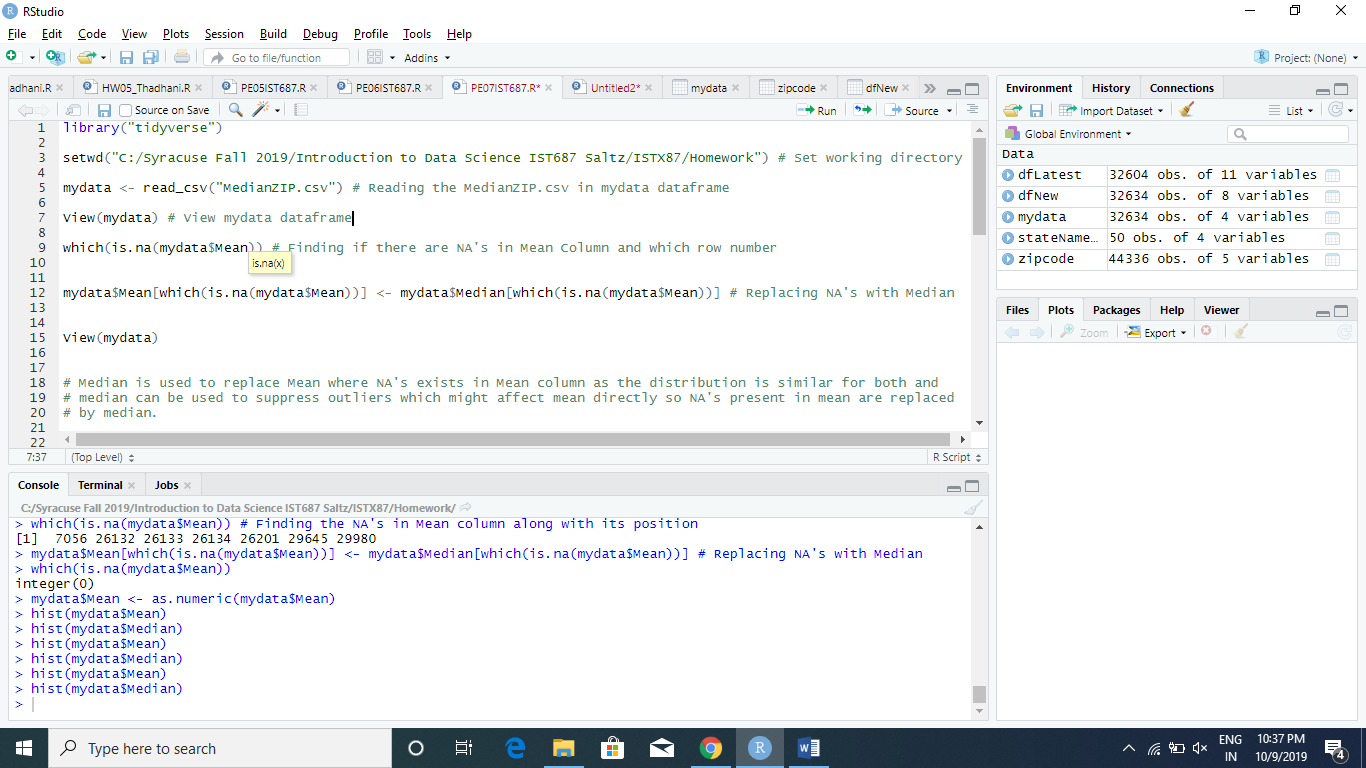
1. For this exercise you will answer all of the questions in this document and turn it in to Blackboard.
2. Before you get started make sure to read Chapters 13 of *An Introduction to Data Science* and execute the code throughout the chapter to gain familiarity.
3. Getting Started: Last week we explored data visualization in R using the ggplot2 package. This week we continue to use ggplot, together with a companion package called ggmap. This companion package enhances the capabilities of ggplot by adding the capability to draw geographic outlines (polygons), shading, labeling, and other map markings.
4. As usual we’ll use the Prep Ex to clean and prepare our data. We’ll use a CSV file that contains information about median incomes. First, we will get our data into a better state so we can use it for the homework. In the homework we’ll again use ggplot and ggmap.

* # IST 687, Standard Homework Heading
* #
* # Student name: Thadhani Hitesh Chandrakumar
* # Homework number: PE07
* # Date due: Wed 9nd Oct 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 Ready: Load and repair median income data**
   1. Download the provided MedianZIP.csv file from Blackboard and read into R-studio into a dataframe called “mydata”. ***HINT: Use read\_csv() to simplify later steps!***
   2. **Cleaning up the NAs:** Find and fix the missing data in the Mean column by substituting the value from the Median column in place of the missing mean values. Explain why the median is a reasonable replacement for the mean.

Median is used to replace Mean where NA's exists in Mean column as the distribution is similar for both and median can be used to suppress outliers which might affect mean directly so NA's present in mean are replaced by median.



* 1. Examine the data with View( ) and add comments explaining what each column contains. Add a comment explaining why the first 2391 zip codes look weird.

The Zip Column contains the zip codes of the various states in USA.

The Mean is the mean of the population.

The Median is the median of the population.

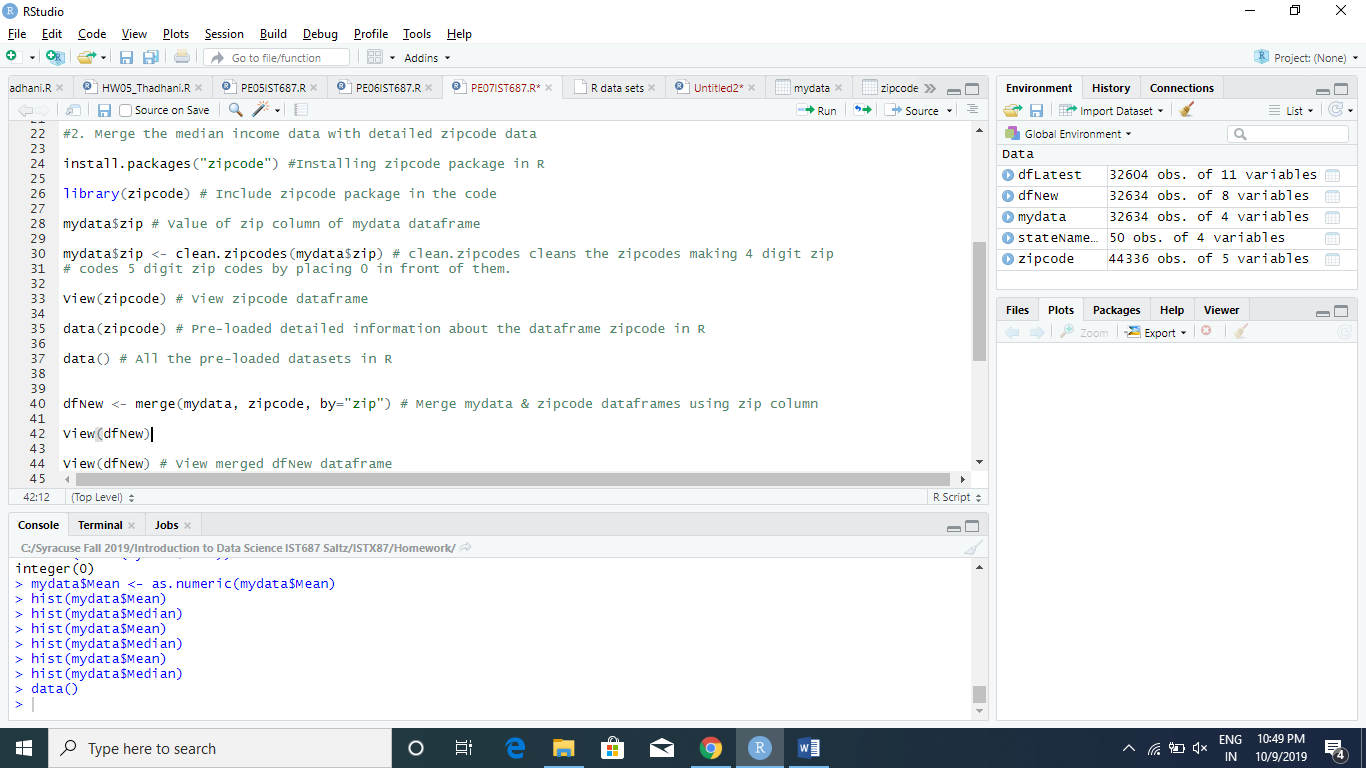
The Pop is the total population of that particular zip code area.

The first 2391 zip codes have only 4 digits in them which is odd as there are 5 digits/numbers in a zip code. So the 1st 2391 zip codes should be 5 digits.

1. **Merge the median income data with detailed zipcode data**
   1. Code and execute the commands below. Write a paragraph below that explains what this code does.

install.packages("zipcode") # Installs zipcode package in R  
library(zipcode) # Include zipcode package in the code  
mydata$zip <- clean.zipcodes(mydata$zip) # clean.zipcodes cleans the zipcodes making 4 digit zip codes 5 digit zip codes by placing 0 in front of them.  
data(zipcode) # Pre-loaded dataframe detailed information in R  
dfNew <- merge(mydata, zipcode, by="zip") # Merge mydata & zipcode dataframes using zip column.

**The Code is padding 0 to the start of zipcodes to make it 5 digits rather than 4 digits as given.**



1. **Merge the new dataset with stateNameDF data**
   1. Create a new dataframe with the following code:

stateNameDF <- data.frame(state=state.abb, stateName=state.name, center=state.center)

stateNameDF$stateName <- tolower(stateNameDF$stateName)

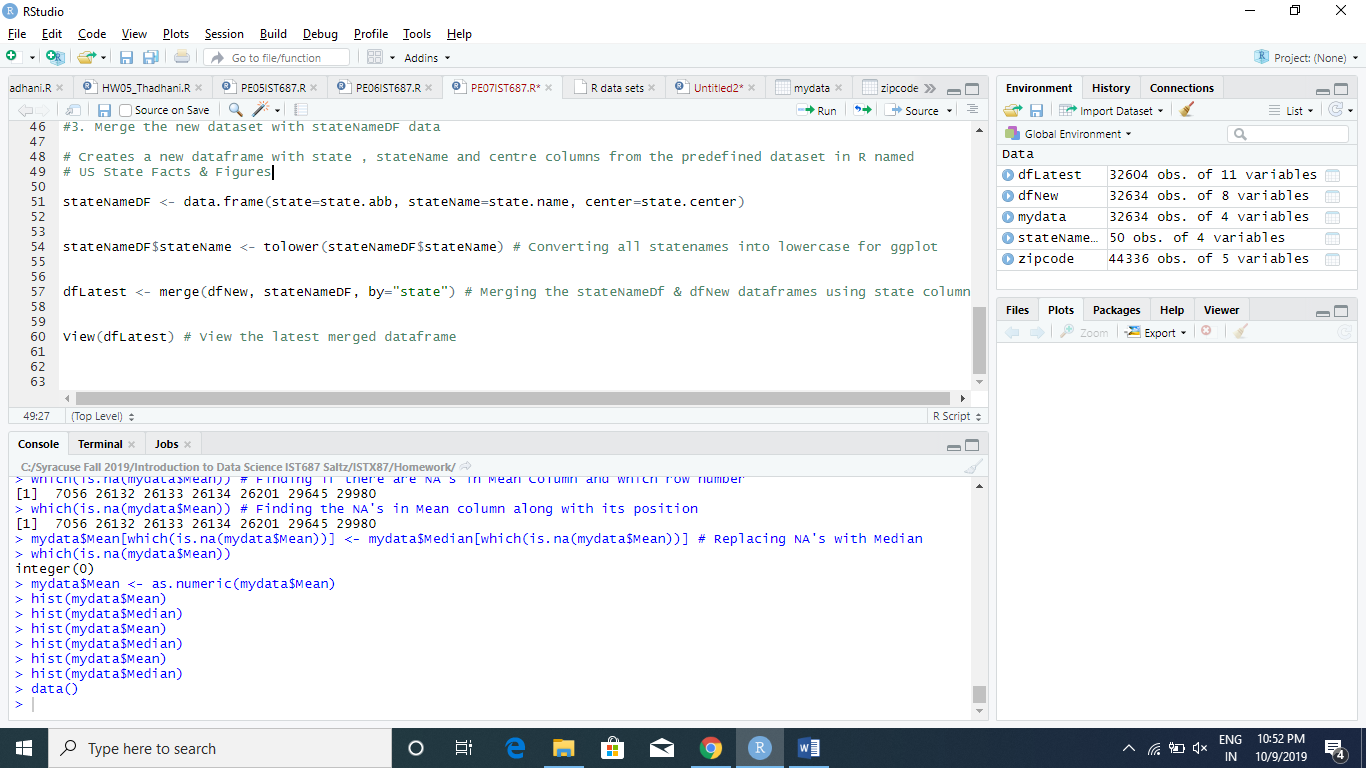
* 1. Comment each line of the code to explain what it is doing

# 1. Creates a new dataframe with state , stateName and centre columns from the predefined dataset in R named US State Facts & Figures

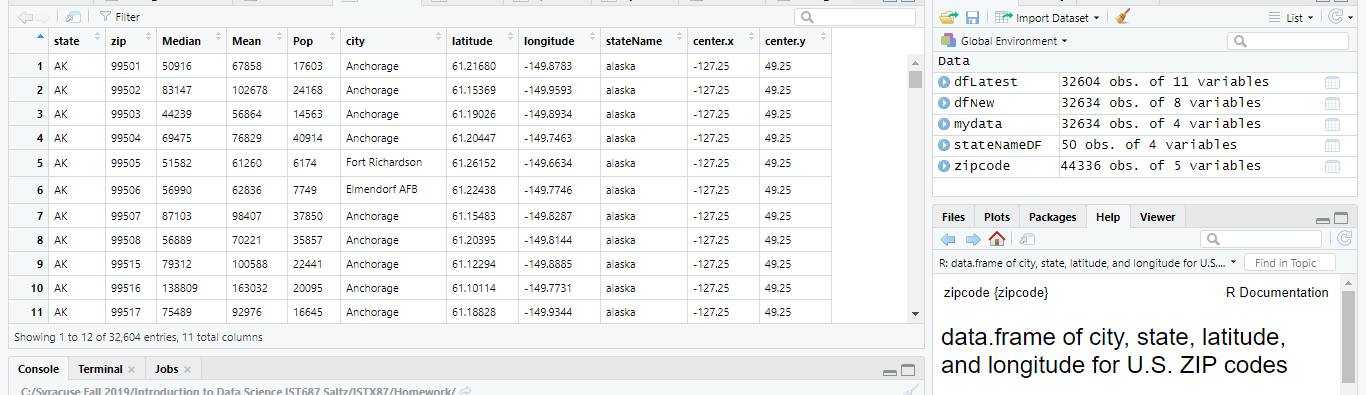
#2. Converting all statenames into lowercase for ggplot as it accepts lower case letters

* 1. Using steps similar to step 2 create a new dataframe that contains our previous information and the information from the stateNameDF.

dfLatest <- merge(dfNew, stateNameDF, by="state") # Merging the 2 dataframes using state column



1. **Examine your new df with the View command.** I
   1. Include a screen shot of the first 10 rows of data and all of your columns.



1. **List any additional resources you used here.**
2. **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.*