# 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.

### 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.

**There are no NAs (missing values) in the Mean column of median income data.**

**Median is the middle value in a sorted list. In case of a symmetric distribution mean=median and for an asymmetric distribution, mean lies closer to the median. Hence, median is a reasonable replacement for the mean.**

* 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.

**#zip: ZIP code of an area**

**#Median: median income in the area**

**#Mean: mean income in the area**

**#Pop: population in the area**

**#the first 2391 observations look weird as they have a 4-digit zip**

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") **#installing zipcode package**   
library(zipcode) **#librarying the zipcode package**   
mydata$zip <- clean.zipcodes(mydata$zip) **#cleaning up zip codes by prefixing 0 to zip where zip is not 5-digit**   
data(zipcode) **#loading the zipcode dataframe**   
dfNew <- merge(mydata, zipcode, by="zip") **#merging mydata and zipcode dataframes by zip into a new dataframe dfNew**

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)

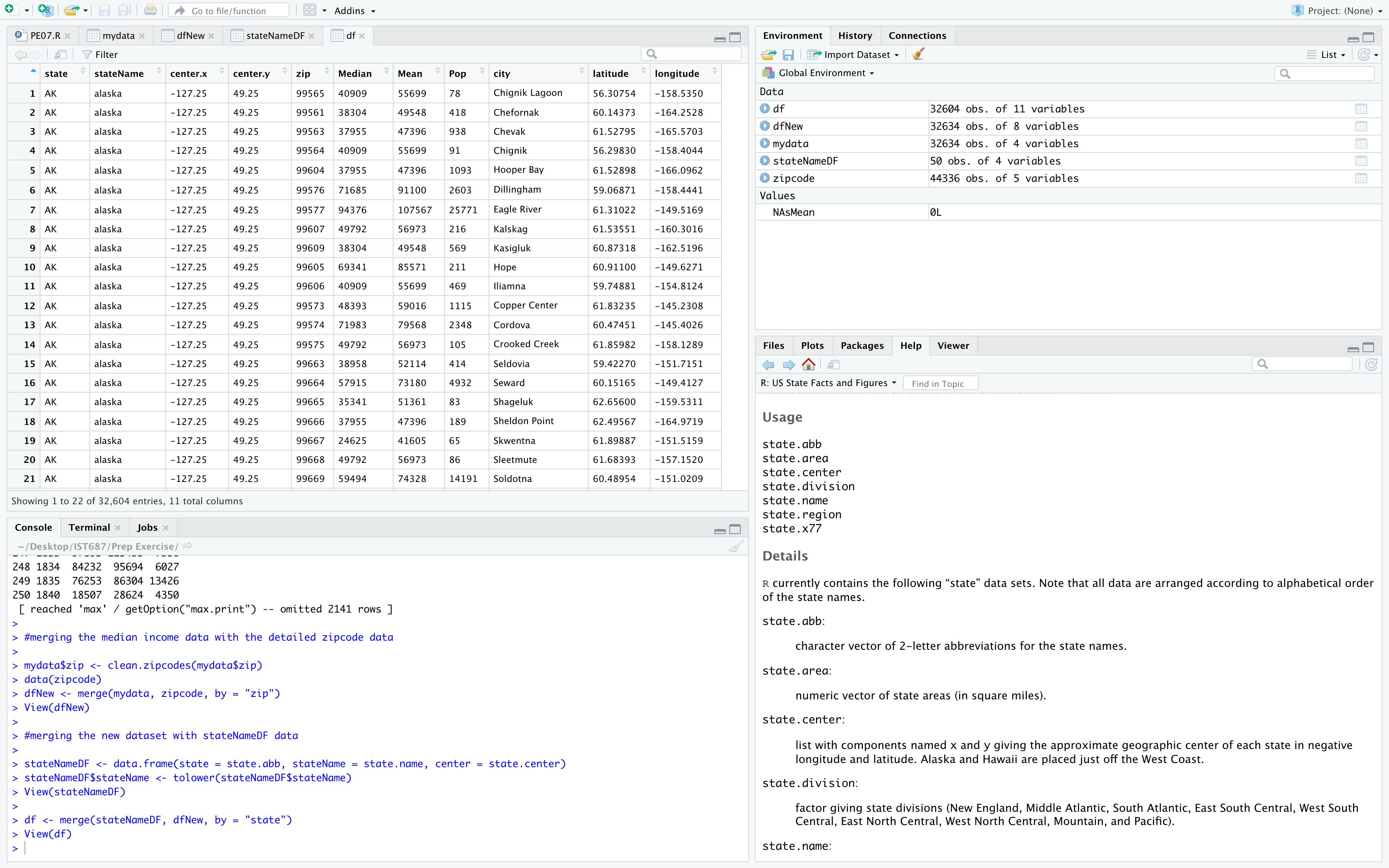
* 1. Comment each line of the code to explain what it is doing

stateNameDF <- data.frame(state=state.abb, stateName=state.name, center=state.center) **#creating a new dataframe stateNameDF for 50 US states with #state attribute as 2-letter abbreviation of state name #stateName attribute as full state name #and center attribute as negative longitude(x) and latitude(y) of geographic center for each state** stateNameDF$stateName <- tolower(stateNameDF$stateName) **# converting stateName observations to lower case**

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

**df <- merge(stateNameDF, dfNew, by=”state”) #merging stateNameDF and dfNew dataframes by state into a new dataframe df**

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.**

RStudio help

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.*