Stat 604
Assignment 05 R

Scope:

This assignment reinforces the material covered in R Lessons 1 through 5, primarily the text handling functions.

Specific Instructions for this Assignment:

There is a file named zip_codes.csv in the **Assignment Data Files** section in eCampus. This text file contains information on all the zip codes for the United States and its territories. Download this file to the folder where you are storing R data on your computer. CAUTION: This file is over 5 Mb in size. If this size prevents you from downloading the file, a zip archive named zip_codes.zip, containing a compressed version of the file, is also provided for your convenience. If you download the zip file, you will need to extract the CSV file before using it. View the data so you will have an understanding of what you will be working with. When you have finished viewing the data DO NOT save it if asked to do so by the application you were using to view it. If you do, you risk changing the data to a format this is not compatible with R.

In addition to the NetID_HW05_console.pdf and NetID_HW05_script.pdf like you submitted in the previous assignment, this assignment will require a NetID_HW05.csv file containing some text output from this assignment. This is one of the few times you will not be converting a file to PDF. There are three "questions" on WebAssign; one for each of your homework files from this assignment. Make sure each file is properly uploaded to the appropriate question. If the upload is successful, you should see the name of the file listed as a hyperlink below the browse box.
Perform in R each of the exercises listed below. Include the required header information as defined in the previous assignment. Also include a comment line in your script above the section for each step so that each is clearly identified. Some of the steps will require multiple lines of code to complete.

A PDF file containing portions of the console log has been posted on eCampus for your reference. Except for the time stamp, the data and order of your output should match the sample output provided.
NOTE: Unless specified, you may use object names of your own choosing.

1. At the top of your script, include housekeeping steps to first list the contents of the workspace and then clear it out.
2. Include a line of code to load the HW04 workspace that you downloaded for the previous assignment. We will be using this data frame again for this assignment. Show the contents of the workspace.
3. Create a new data frame from the Oklahoma data. The new data frame will only contain High Schools as indicated by HS somewhere in the school name. Be sure to construct your logic so that you do not pick up schools with JHS in the name. (It is acceptable to include schools that are JR-SR HS.) You will need to look at the data carefully and compare how these records are different as you plan the expression you will use to subset the data. Base your code on a reliable pattern in the data. Use a negative subscript to remove columns that are not shown in the structure of the new data frame in the sample console. Show the structure of your new data frame.

4. Create a new data frame by reading in the zip code database file that you downloaded from eCampus. Show the structure of this data frame.
5. We are considering some analysis on our Oklahoma High School data by Zip Code regions but that data frame does not contain zip codes. We want to bring in zip codes from the database that we downloaded but there are some limitations we must deal with and will still be limited in what we can do. Prepare the zip code data for combining with the high school data:
    a. Create a new data frame of Oklahoma zips based on the value of state. Eliminate rows with decommissioned zip codes and PO BOX type zip codes. Look at the values in the decommissioned column and consider how we use values 0 and 1 as you consider how to construct your logic. Include only the zip, primary_city, county, and estimated_population columns in the new data frame.
    b. Use the substitute function to change the name of primary_city to MailCity. Remember, you can use the names function to access column names for both reading values and assigning values.
    c. Change the MailCity values to upper case.
    d. One of the problems with this process is that larger cities have a number of zip codes and we do not have enough address information from our schools to identify the specific zip code. This is going to cause multiple records which we will deal with later. Since we do not know the specific zip code, we want to create a zip region based on the first three digits of the zip code. Use the substring function to create a new column of character values that contain the first three digits of the zip code.
    e. Display the structure and the first 20 records of the new Oklahoma zip code data frame.
6. Create a new data frame by merging the Oklahoma High School data with the Oklahoma zip code data. Show the dimensions of the new data frame.
7. The function **duplicated(OKHSzips$School)** can be used to provide a mask of values in which the School name is a duplicate record (Use the name of your data frame). Use this function in an assignment statement to create a data frame of records that are NOT duplicates. Show the structure of this new data frame.
8. Display the 25 smallest and largest schools based on the number of Teachers from the unduplicated data frame. Make sure the rows and columns are in the order shown in the sample data provided on eCampus.
9. Use cat and/or paste to write an R expression that will automatically create a CSV text file of schools from the unduplicated data frame. Include the School, MailCity, County, ZipRegion and HSTotal columns in that order. Add another column that contains the system date and time. Except for the time stamp, your data should match the data in the screen shot below that shows the first few records of the file. Include the file path in your script code. NOTE: The write.table function would be the preferred method for creating a CSV file but we are using this method to give you an opportunity to practice using the cat and paste functions.

```
LATTA HS,ADA,PONTOTOC COUNTY,748,149,2016-09-10 20:30:37
ADA HS,ADA,PONTOTOC COUNTY,748,502,2016-09-10 20:30:37
VANOSS HS,ADA,PONTOTOC COUNTY,748,171,2016-09-10 20:30:37
BYNG HS,ADA,PONTOTOC COUNTY,748,302,2016-09-10 20:30:37
ADAIR HS,ADAIR,MAYES COUNTY,743,285,2016-09-10 20:30:37
AFTON HS,AFTON,OTTAWA COUNTY,743,121,2016-09-10 20:30:37
AGRA HS,AGRA,LINCOLN COUNTY,748,98,2016-09-10 20:30:37
ALEX HS,ALEX,GRADY COUNTY,730,106,2016-09-10 20:30:37
ALINE-CLEO HS,ALINE,ALFALFA COUNTY,737,38,2016-09-10 20:30:37
ALLEN HS,ALLEN,PONTOTOC COUNTY,748,128,2016-09-10 20:30:37
ALTUS HS,ALTUS,JACKSON COUNTY,735,1038,2016-09-10 20:30:37
NAVAJO HS,ALTUS,JACKSON COUNTY,735,96,2016-09-10 20:30:37
```