CSE 411: ADVANCED PROGRAMMING TECHNIQUES

Fall 2016

R Programming Homework

Sachin Joshi

Program Description

The R file **R_HW.R** contains the code that accomplishes the following two tasks:

Task 1: Takes the Census data and plots the relationship between 2013 population size and education (percentage of people with a bachelor's degree) for the counties in each state to see if that relationship doesn't hold in some states but does in others.

Task 2: Uses the **data ( )** command to review the data sets available in the installed packages that can be combined in some useful way with R and generates a chart showing some aspect of that combination.

Running the R file

Before running the actual program make sure that all the datasets namely **DataDict.txt**, **DataSet.txt** and **FIPS_CountyName.txt** as well as the executable R file **R_HW.R** are contained within the same folder. To execute our program we must follow the below mentioned steps:

- Download the package **R_Prog_HW_saj415.zip** and unzip all the files into a folder in your sunlab machine.
- Launch the **RStudio**.
- Set the console path to the current directory that contains all the unzipped files by running the following command in RStudio:

  setwd("PathToDownloadedFilesFolder")
- Execute the R file **R_HW.R**.

Output File

The pdf file **R_Prog_HW.pdf** contains the results of both of the above mentioned tasks, Task 1 and Task 2.

1. **Task 1:** Plot of the relationship between 2013 population size and education (percentage of people with a bachelor's degree) for the counties in each state.
2. **Task 2:** A chart showing some aspect of the combined data sets available in the installed packages.

Program Description

The code has been split into two different sections to achieve the objectives of both the tasks, Task 1 and Task 2.

Task 1: To achieve the desired results, Task 1 follows the below mentioned steps:

1. **Data Accumulation:** Collects the data from the different data sets and merges it into a single data frame variable.
2. **Data Pre-processing:** Pre-processes the data to a usable dataset for the end requirement. Gets all the state names, then uses the state names to convert all the counties in their corresponding states. Also removes the unnecessary data like the row containing the country name "United States". Finally creates a duplicate dataset so that it doesn't affect the original data set.
3. **Data Visualization:** The final step is to visualize the collected data and plot it into different subplots. The end product can be seen in the output PDF file, **R_Prog_HW.pdf**.

Task 2: To achieve the desired results, Task 1 follows the below mentioned steps:

1. **Data Accumulation:** This step gets the **UKDriverDeaths** and **UKgas** details. We want to find the correlation between these two data sets and eventually you will notice that as the gas usage keeps increasing every year, the amount of deaths start decreasing. We can see that, more and more people are getting cars every year and people are becoming more careful while driving.
2. **Data Pre-processing:** Converts the quarterly data into yearly data for both the datasets and then combines them into a single data frame variable.
3. **Data Visualization:** Uses ggplot to plot the relationship between both these data sets into a single graph. The end product can be seen in the output PDF file, **R_Prog_HW.pdf**.