CMPSC 380
Principles of Database Systems
Fall 2014

Laboratory Assignment Two: Procedural Programming and File Processing Systems

Introduction

In this laboratory assignment, we will use the procedural (or, imperative) approach to implementing a simple file processing system. In this assignment you will familiarize yourself with the steps that a scientist would take to analyze, manipulate, and visualize a data set. In particular, you will learn how to use the R language for statistical computation to manage and visualize a file-based data set. You will also develop a preliminary understanding of how how to write simple procedures that select a subset of data from a larger data set. Also, you will try to summarize data sets using functions such as the mean and median. Once you have a better understanding the challenges associated with the use of imperative programming techniques during data analysis and management, you will also investigate the steps needed to produce visualizations of a file-based data set.

Installing and Configuring the Data Sets

During the completion of this laboratory assignment, we will rely upon some of the tools and data sets provided by Luis Torgo's book entitled "Data Mining with R: Learning with Case Studies". A copy of the book will be available for your consultation throughout the laboratory session and then held on reserve in my outer office. You can also learn more about this book by visiting the following Web site: `http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR/`.

As previously mentioned, this assignment invites you to implement procedural (or, imperative) methods—in the R language for statistical computation—to answer questions about a real-world data set. To start this laboratory assignment, you should start using the command-line-based and interactive environment of the R language. You can do this by typing "`R`" into your terminal window. After making sure that you already have a Web browser open, you should next type the command "`help.start()`" in order to load the help system for R. You can use this Web-based help environment throughout the laboratory session as you learn new commands for imperative data management. In addition, you can learn more about R by visiting Web sites such as `http://www.statmethods.net/` and `http://www.cyclismo.org/tutorial/R/`.

To access the data files needed for this assignment, you must install the software package for the companion textbook. Since this process may takes a long time, you should type the following command in your R shell right away: "`install.packages("DMwR")`". Once you have pressed enter, you will see that many lines scroll in your terminal window as R performs a network install of all the required packages, including the needed data files. When this process finishes, you can load the library by typing "`library(DMwR)`". You should also repeat these steps for the "`Hmisc`" package.

In order to effectively use R to analyze the data in a file, you need to learn several basic commands. You can view the manual for a specific command, say `subset`, by typing "`?subset`" at the R prompt. In this phase of the assignment, you should learn how to use several R commands and write a short description of their input, output, and behavior, including one concrete example of the command in action with the data set discussed at a later stage of the assignment. Please see the course instructor if you have a question about how to use a data manipulation command in R. In particular, you must study the following commands available in the R language:

- `attach`
- `names`
- `head`
- `mean`
- `median`
- `subset`
- `ls`
- `summarize` or `summary`

Procedural Exploration of a Data Set

After you have learned more about how each one of the aforementioned commands works, you should learn more about the `algae` data set that is part of the `DMwR` library. What are the names of the attributes in this data set? How many attributes are in the data set? How many rows are in the data set? While you must use the R programming language to answer these questions, you can check your responses by visiting the "Algae Data Set Description" in the UCI Machine Learning Repository and reading the "Predicting Algae Blooms" chapter in Torgo's book.

As you learn more about this data set, you will notice that the `a1` attribute gives the frequency number of a harmful algae known as "a1" in the data set. Please note that the data set does not contain any information about the name or the characteristics of this type of alga. Each value in the `a1` attribute corresponds to a "frequency" that this algae was found in the specified environment; in this case, small values are better since they indicate the presence of less of this harmful algae. Using the mean and the median values of the frequencies for alga `a1`, is this algae more likely to bloom in rivers classified as "small", "medium", or "large"?

The `a2` and `a3` attributes respectively give the frequency number of a harmful algae known as "a2" and "a3" in the data set. Using the mean and the median values of the frequencies for algae `a2` and `a3`, are these algae more likely to bloom in "small", "medium", or "large" rivers?

Additionally, the `algae` data set also contains details about the speed of the rivers, as stored in the `speed` attribute. Using the mean and the median values for the "a1", "a2", and "a3" algae, are these algae more likely to bloom in rivers with a "low", "medium", or "high" speed?

The `Cl` attribute in the `algae` data set describes the mean amount of chlorophyll in the rivers. What is the relationship between the mean value of chlorophyll and the amount of "a1", "a2", and "a3" in the rivers? For instance, if a river has a "high" amount of chlorophyll, does this mean that it will contain a "high" or a "low" amount of each algae? Why do you think this is the case?

There are many other interesting trends evident in the algae data set. Using any combination of R commands, please identify and try to explain one additional trend. As you are completing this last part of the laboratory assignment, it may be helpful to visualize the `algae` data set so that you can easily and quickly spot new and interesting phenomena. To create a wide variety of different graphs, you can explore the use of the "Lattice" package that is available after you type the "`library(lattice)`" command. You can learn more about Lattice by visiting the package's Web site that is available at `http://lmdvr.r-forge.r-project.org/`. Once you have learned how to use Lattice to produce simple data visualizations in R, you should construct at least one graph to accompany your response to one of the previously stated questions.

Summary of the Required Deliverables

This assignment invites you to submit one printed version of the following deliverables:

1. A screenshot showing a small selection of the output seen when installing "`DMwR`".
2. The description of the input, output, and behavior for the required R commands.
3. Answers to all of the questions posed, with supporting output and evidence as appropriate.
4. At least one supporting visualization of some trend that you found in the data set.
5. A commentary on the challenges that you faced and the way(s) that you overcame them.

In adherence to the Honor Code, students should complete this assignment on an individual basis. While it is appropriate for students in this class to have high-level conversations about the assignment, it is necessary to distinguish carefully between the student who discusses the principles underlying a problem with others and the student who produces assignments that are identical to, or merely variations on, someone else's work. Deliverables that are nearly identical to the work of others will be taken as evidence of violating Allegheny College's Honor Code.