DA 320 Assignment 10 M. Blanco

**Part 1: Short Answers (use 3-5 lines to answer the following questions)**

**Problem 1:** Discussion the use of R and what it was originally designed to do.

R is a interpreted language and environment designed for statistical computing (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc), data manipulation, calculation and graphical display. R can be extended through packages.

**Problem 2:** Explain what a data frame is and discuss why some data can’t be easily expressed as a collection of similarly typed numbers.

An R data frame is used for storing data tables (two-dimensional array-like structure). Data frame is a list of vectors of equal length. Column contains values of one variable and row contains one set of values from each column.

Data frame have the following characteristics:

* The column names cannot be empty.
* The row names should be unique.
* Data frame can store numeric, factor or character data type.
* Each column contain same number of data items (row).

**Problem 3:** Discuss a common fallacy of large data analysis problems, and give a detailed real life example.

A common fallacy is bigger data not always better data.

Recently Harvard Business Review conducted an experiment by sending 316 fake applications to the largest U.S. law firms. Those that received the application were in the top 1 percent of students at their schools. In this case Harvard Business Review study concluded that using less data would have been a better option.

https://www.bloomberg.com/view/articles/2017-03-01/bigger-data-isn-t-always-better-data

**Problem 4:** Explain why it is considered that data frames are the workhorse data structures of the R language.

A data frame is the most common way of storing data in R. Data frame makes it easy to analyze data.  Data frame is a collection of vectors that all have the same length. Data frame are similar to matrix, except that each column can contain a different data type. A data frame is used to represent an entire data set.

**Problem 5:** Explain when and why regression analysis is easy to misuse and why it is easy to misinterpret the results.

There are three common abuses of regression analysis:

* Extrapolation
  + Extrapolation outside the model data. It important to infer the domain of independent variables for which a given equation is valid.
* Generalization
  + Occurs when unsupported or overexaggerated claims are made
* Causation
  + Regression analysis can only help confirm or refute a causal model

**Part 2: Research Analysis and Case Study**

**Problem 6:** On page 150 of the book the author explains the concept of bigmemory and in what situations it is used. The author gives the example of the Department of Transportation’s airline on-time database and provides a snippet of sample code in Listing 11.3 showing how bigmemory is used. Study this database and determine what other databases it bigmemory would be a useful tool for analyzing data, and how? Use application examples(s) to explain in details.

Bigmemory can be used to analyze Behavioral Risk Factor Surveillance System (BRFSS) dataset. BRFSS is the national health-related telephone survey in all 50 states as well as the District of Columbia and three U.S. territories. Subjects are asked questions regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. BRFSS completes more than 400,000 adult interviews each year. BRFSS is the largest continuously conducted health survey system in the world. The survey was established in 1984.

library(bigmemory)

brfss <- read.big.matrix(“brfss.csv”, header=True, backingfile=”allbrfss.bin”, descriptor=”allbrfss.desc”)

head(brfss)

summary(brfss)

**Problem 7:** In the article [http://www.stat.yale.edu/~mjk56/temp/bigmemory-vignette.pdf,](http://www.stat.yale.edu/~mjk56/temp/bigmemory-vignette.pdf) the authors gave an example of Netflix Prize data. Write a short summary on this article (no more than one page). In what scenario would bigmemory be a useful tool for analyzing data? Use application example(s) to explain in details.

R was not designed to explore large dataset. Both data frames and matrices were not designed for large dataset. The bigmemory package can be used to read and analyze massive data set. This package exploit the available RAM on a computer to analyze data. Bigmemory extends R which allows for more powerful parallel analysis and data mining on large data sets.

Bigmemory provides a matrix-like functionality for large data set that could be larger than a computer’s available RAM. Bigmemory provides data structure that hold the entire data set. There is no need to load and unload data. The data can be accessed and manipulated similar to R matrix object. Last, bigmemory works with many R standard functions. Bigmemory was designed for parallel and distributed computing enviorment.

Example 1: Review SAT data

**library**(bigmemory)

school.matrix <- read.big.matrix(

"SAT.csv",

type ="integer", header = TRUE, backingfile = "school.bin",

descriptorfile ="school.desc", extraCols =NULL)

**library**(ff)

school.ff <- read.csv.ffdf(file="SAT.csv")

Example 2: Researchers perform spatial analysis using bigmemory and ff.

Files used:

Satellite scene somewhere in Germany

<http://image2000.jrc.ec.europa.eu/>

File de33\_194025.bsq

.bsq filesize 334 MB

Satellite scene somewhere in Ethiopia

<http://www.orthocoverage.com/download/> (This link did not contain the data file)

File DEMO\_ETHIOPIA\_1\_IMAGES\_SWIR\_1.tif

tif filesize 11 MB

References:

http://nuest.staff.ifgi.de/pub/bredel-nuest\_handling-large-image-data-in-R.pdf

https://rstudio-pubs-static.s3.amazonaws.com/72295\_692737b667614d369bd87cb0f51c9a4b.html

https://www.youtube.com/watch?v=59HEplUji\_I&feature=youtu.be