CSE487 Data-intensive Computing Spring 2016

Due date: 3/5/2016 by 11.59PM or earlier, online submission

Title: Problem solving and exploratory data analysis using R.

Introduction and Context: Data Science process involves a phase where preliminary understanding of the data is explored. This phase is commonly known as exploratory data analysis (EDA) [1, 2]. This is a very useful phase before launching onto a full scale big-data analytics that often requires extensive infrastructure support and complex or newer algorithms such as Hadoop MapReduce [3] and Apache Spark ML [4] on large clusters. The algorithms and methods we will use in EDA are also highly applicable and useful to perform (run) statistical analysis on commonly available public data sets and on the outputs of big-data processing. In my experience with data, EDA is perfect complement to big-data analytics. EDA lets you gain intuition about the data in the early phases and it also offers you a systematic approach to choosing among the many possible outcomes of big-data analytics. EDA begins with understanding the schema of the data (head) and ends with an approximate statistical model representing the data.

Problem Statement: You will work on a fairly large data sets. The data you choose should have been collected for a period of time. The example in the book for NY Times is for a month. You will run statistics on this data to perform EDA using the process and steps listed in Chapter 2 of your text book. You are required to prepare a report document based on the EDA. The report also should provide the provenance for the results and outcomes (graphs etc.) listed in the document.

Data Characteristics: We will use structured and semi-structured data we mean data stored in tables (files), databases (RDBMS) and spreadsheets (like Excel CSV files). The entire data could be stored in several files in tables, and spread over multiple directories (depending on the data). Download and extract the complete data set provided with your text book. It is available at
https://github.com/oreillymedia/doing_data_science

We describe the problems to be solved in several parts.

Problem 1: (20%) Data acquisition: Data collection is an important step that often is quite time consuming and constrained by regulations and privacy and security issues. For instance in order to collect data from patients in a clinical trial, a researcher in an organization has to get prior approval and certification from Internal Review Board (IRB) of the organization. There are many regulatory laws governing what data you collect about common citizens. **Students ought to be sensitive to all these when collecting data.** The other challenge in data collection is that data comes in different formats (text, html, txt, csv, json etc.) and feature widely varying access methods (web URL, api, hdfs, sqldb etc.). We will learn about this by working on few representative methods for data acquisition given in the handout: http://www.cse.buffalo.edu/~bina/cse487/spring2016/Lectures/RHandout1.pdf

To that list of methods add an approach for reading json data. For example twitter data is published as JSON objects.

Problem 2: Simple EDA (20%) This problem explores the effectiveness of online newspaper promotions and advertisements. New York Times data collected in chapter 2 contains these information about readers: {age, gender, number of impressions, number of clicks, and logged in or not.} The EDA process from loading the data to plotting charts is explained in the sample code in pages 38-40. (a) Understand these steps by executing the sample code on R and RStudio. Save the scripts, plots, and data environments (for a possible TA demo). Name the R script NYTP2Username.R. (b) Now extend the EDA to monthly data and follow the questions given in Page 38. Save the script as NYTP2ExUsername.R. (c) For both problems save the charts and interpret the outcomes. Write a paragraph to interpret the results. Save this in a single pdf document NYTP2Username.pdf. zip or tar and submit NYTP2Username.tar

Problem 3: Data economy: A *real* case study (20%). For this problem you will work the RealDirect problem discussed in Ch.2. Work on the sample code discussed in p.49,50. Then extend it to the entire RealDirect data set to perform EDA to find some more insights. **You will realize the data given as well as the code require some cleaning and editing.** This particular problem is a "data problem" from a company is that is currently operational, see http://www.realdirect.com/ . Read the details and understand the **business model**. The sample code is given in pages 49-50 for Brooklyn borough. Work on it and make sure understand the EDA process. Then answer the questions in pages 48-49 in your report. Repeat the analysis for different boroughs. Save the scripts, plots, **matrix of plots to compare** various boroughs, and data environments (for a possible TA demo). (a) Name the R script RD3Username.R. (b) Now extend the EDA to monthly data and work on the questions given in page 48-49. Save the script as RDP3ExUsername.R. (c) For both problems save the charts and interpret the outcomes (charts): write a paragraph to interpret the results. Save this in a single pdf document RDP3Username.pdf. zip and tar and submit RDP3Username.tar

Problem 4: Statistical analysis to support new data product (20 %) In this problem we will further explore the RealDirect business. You should realize by now RealDirect has built a business around existing real estate (buying and selling) business by creatively repurposing the data and building a ***data product*** around it. They have created a web (and mobile) portal with tools to facilitate real estate related operations. Assume you have been hired by ReaDirect to extend the line of product offerings. You put on your thinking cap and realize that NY is a prime location for ***apartment rental*** since buying real estate (houses and apartments) is beyond their means for many. You also realize that many prospective clients take to twitter when they need something and want to express their sentiments and status. You plan to recommend to the executive team at RealDirect that they should offer apartment rental as a product. You want to arm yourself with data to prove your recommendation. You plan to collect twitter data about apartment rental and real estate (buying a house) for a week on a daily basis and show the feasibility of your recommendation with this statistical analysis. Also provide a pricing model (ex: subscription or one time registration etc.). Prepare the tar file as suggested in Problems 3 and 4 and submit the RDExP4UserName.tar

Problem 5: Stream processing (20%): Processing and analyzing streaming data is an important emerging area. A stream is a (digitized) sequence of events. (Streaming applications are applications that are deployed on demand: we will focus on that in Project 2.) Example of streams include sensor data, news events and tweets). It is election season in the USA now, why not observe who is trending and do some daily and weekly summarization. We will collect tweets from various regions of the country and plot the daily and weekly trend and summarize the stats. These trends and summarization are to be displayed on web /mobile portal created by **R Shiny** that is a web application framework for R. See http://shiny.rstudio.com/ and study the examples. Shiny has a web client front end powered by R on the server end. You can really create a very nice election dashboard with live streaming data. For inspiration look at the examples of Shiny dashboards at http://shiny.rstudio.com/gallery/ . You can use data any other domain too (Eg. Financial markets and stock data).

Implementation details:

1. **You work on your own (group work is NOT allowed in this project.)** You can source your data from multiple sources while this is NOT a requirement. You may have to clean the data.
2. Now for the most critical aspect: What statistics will you run on the data collected? The statistics discussed in Chapter 2 is quite basic and you should be able to run summaries, plots and pdfs as discussed in Chapter 2. Perform these and make sure you record the outcome in your report.
3. Perform the analysis for a single data set (file). Extend it to multiple files/data sets. Visualize some metrics and distribution over time. Describe all the outcomes in your report.
4. You are required to prepare a detailed report and documentation that will allow us to repeat the experiment/analysis you have carried out and also provide the provenance for the results you have generated.
5. Use elegant directory structure and naming conventions for directories and files to capture all the work for project 1 and then tar or zip each problem into its own compressed file of self-explaining names and problem# and username.
6. Submit the solutions as soon as you complete them. Do not wait till the last minute to work on all the parts.
7. **Recommended** dates for completion of the parts proving the feasibility of the project completion within the allocated time: 2/12, 2/14, 2/21, 2/28, 3/5
8. **Do not publicize the code on github or any other similar public forum until the semester is over**.

References:

1. C. ONeil and R. Schutt. Doing Data Science. Orielly, 2013.
2. J. Tukey. Exploratory Data Analysis. Pearson, 1977.
3. Hadoop-mapreduce. http://hadoop.apache.org/, last viewed Jan 2016.
4. Spark ML, https://spark.apache.org/docs/1.2.1/ml-guide.html, last viewed Feb 2016.