Fall 2016
MIS 413/572 - Introduction to Big Data Analytics

Homework 1

Graded out of 100 points. Due in class on December 2nd. Please typeset your homework, save as an R source code file with the title of "your student ID-Homework_1.R" (e.g. B1234567-Homework_1.R), and submit to NSYSU Cyber University. Note that your code must follow the suggested programming and data analysis styles discussed in the class.

1. Please download loan datasets 2007-2015 (“LoanStats3a-d.csv”) with the data dictionary on LendingClub.com.

1.1 **[5 pts]** Please load all loan datasets. Skip those records after row number #39786, #188183, #235631, and #421097 respectively (records with “Loans that do not meet the credit policy”).

1.2 **[5 pts]** Concatenate these datasets into an R data frame, and only keep those with *loan_status* in “Fully Paid” and ”Charged Off”.

1.3 **[5 pts]** Remove those columns with any NAs.

1.4 [**5 pts]** What is the percentage of the “Charged Off” loan?

1.5 **[10 pts]** Please replace below R code with SQL code that does similar split-apply-combine operations. Suppose “loan” is the name of your concatenated data frame.

```
# Split, by emp_length
sp_loan = split(loan, loan$emp_length)
# Apply, get average loan amounts
result = sapply(sp_loan, function(x) mean(as.numeric(x$loan_amnt)))
# Combine, into a data frame
result = data.frame("Employment_Length" = names(result),
  "Loan_amount_average" = unname(result)); result
```

1.6 **[10 pts]** Please replace below SQL code with R code that does similar data management tasks. Suppose “loan” is the name of your concatenated data frame.

```
# For those of top (> 5000) loan purposes,
# count the number of loans for different grades
SELECT grade, count(id) as Grade_Count
FROM loan WHERE purpose IN
      (SELECT purpose FROM loan GROUP BY purpose HAVING count(id) >= 5000)
GROUP BY grade
```

1.7 **[15 pts]** Please upload your loan R data frame to the HDFS of our server. Replace the following SQL code that calculates the average annual income (*annual_inc*) for each *grade* with a MapReduce function. (Hint: test your MapReduce function with a small dataset and make sure it works before applying it to the loan data)

```
SELECT grade, avg(annual_inc) FROM loan GROUP BY grade
```

2. **[15 pts]** Please write a MapReduce function that remove records/observations with any NAs. Assume that the input data format is a native R data frame. (Hint: consider your function a MapReduce version of *stats::na.omit*())

3 . The use of *Closure* introduces a concept of "function factory" in most functional programming languages (e.g. R and Scala). Please refer to the introduction of Closure on wikipedia and answer the following questions.

3. 1 **[5 pts]** Briefly explain what a *Closure* is.

3 . 2 **[5 pts]** Create an R function that returns closure *n_percentage(n)* to list top-N percentage elements in a given numeric vector. For example, the following closure *ten_percentage*() returns top 10% biggest values in a numeric vector.

```
> ten_percentage = n_percentage(10)
> ten_percentage(1:100)
[1] 100 99 98 97 96 95 94 93 92 91
> set.seed(1); ten_percentage(runif(30, 0, 1) )
[1] 0.9919061 0.9446753 0.9347052
```

4. **[20 pts]** Please read the article "Chocolate Consumption, Cognitive Function, and Nobel Laureates" on our reading list. Try to reproduce the correlation analysis but use alcohol consumption instead of chocolate consumption. Specifically, we consider the relationship between the alcohol consumption per capita and all Nobel prizes per capita. Do you see anything interesting? Justify your findings with your analysis results.