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

Exercise 2

1. Please download loan data 2007-2011 (“LoanStats3a.csv”) with the data dictionary on LendingClub.com. Consider the following data management questions.

   1) Import & read the first 39,786 observations of the CSV file for those that “meet the credit policy”. Keep those records with *loan_status* in (“Fully Paid”, “Charged Off”).

   2) Create an R function *procFreq*(*df, x*, *y*) that returns a crosstab of *x* by *y*, and chi-squared statistics/p-values for test of independence between *x* and *y* given an R data frame *df*.

   3) Perform a series of bivariate analyses on loan_status (as the outcome) by *grade*, *purpose* and *term*. What variable(s) might affect the *loan_status*? Use your own *procFreq*() if you’d like.

2. Please install package *vcd* and load dataset *vcd::Arthritis*.

   1) What are the average ages for different treatment (*Treatment*) and the status of improved (*Improved*)? Report your result with a crosstab of *Treatment by Improved*, where each cell contains the average age. Hint: *dcast*() may help.

   2) Use any statistical tests to examine whether *Treatment* and *Sex* are independent.

3. Load the given dataset “cars.csv”.

   1) Keep those records with *fuelType* = “gas” and clean the data by removing those incomplete cases (record with any missing values).

   2) Consider the following SQL code.

      ```
      Select bodyStyle, avg(highwayMpg) as avgHwMPG
      from cars group by bodyStyle order by avgHwMPG
      ```

      Replace the code with R data aggregation functions, or your own split-apply-combine statements.

4. Please convert built-in dataset "Titanic" into an R data frame. Consider the following SQL/R codes.

   1) Replace below SQL code with equivalent R data aggregation functions.

      ```
      Select Sex , Survived , SUM(Freq) from Titanic
      ```

```
where survived = 'Yes' group by Sex, Survived
```

2) Replace below R code with equivalent SQL code.

```
subset(aggregate(x=df_Titanic$Freq,
  by=list("Class"=df_Titanic$Class),FUN=sum),x>300)
```

5. Please login to our RStudio Server on http://hadoop.cm.nsysu.edu.tw:8787. Consider a data management task that we would like common row(s) among 1,000 CSV datasets in "/home/temp/CSVs". Note that you should use R functionals discussed in the class (e.g. *Map*(), *Reduce*(), or apply-family functions), instead of any R loop statements. It should normally take the server a few seconds to execute your R code.