SimpleStatisticalAnalysis

Xiaosong Zhang

9/18/2016

1.What's your proudest achievement and why?

It can be a personal project or something you've worked on professionally. Just a short paragraph is fine.

From my understanding, there are three types of things that I think could be considered as proud achievement:

- Self-improvement: Be able to complete difficult tasks through learning and practice. For example, build a computer from pieces of hardware for the first time; install and run my personal server or improve the performance of the code I wrote, etc.
- Standing out in competition: Prepare and give a presentation that impresses the audiences better than my counterpart; Solve a given problem in a more efficient and elegant way, etc.
- Helping others overcome their obstacles: Tutor students and friends to not only getting understand the questions they cannot figure out but also try to locate the knowledge points they didn't know well which preventing them from figuring out the problem, etc.

2. Tell us about a book or article related to data analysis you read recently, why you liked it, and why we should read it.

3. Tell us about one aspect of the Digital Innovation Greenhouse you really like, and why.

4.STATS250 KEY FACTOR ANALYSIS

Please answer the following question: Other than GPAO, which variable(s) best predict the variable for GRD_PTS_PER_UNIT for the course STATS 250 (SUBJECT="STATS" and CATALOG_NBR=250)? In other words, what is the best predictor of a student's performance in STATS 250 other than the student's own GPA?

Data Import

```
require("dplyr")
require("ggplot2")
#sourceDir is use to source all .r file under same Dir all in once
sourceDir <- function(path, trace = TRUE, ...) {
  for (nm in list.files(path, pattern = "[.][RrSsQq]$")) {
    if(trace) cat(nm,":")
    source(file.path(path, nm), ...)
    if(trace) cat("\n")
  }
}
#Source all functions uder PLA-MOOC into R-Environment for future use.
sourceDir('PLA-MOOC')
```

```r
#loading data into Environment
student.course <- read.csv("PLA-MOOC/student.course.csv")
student.record <- read.csv("PLA-MOOC/student.record.csv")
```

Data Exploration

```r
names(student.record)
names(student.course)

# use summary(student.record) ;summary(student.course) to get a big picture of the data
# HSGPA= 36? it most likely to be a error
# UM admit students with HSGPA 0?? or it should be NAs?
filter(student.record, HSGPA>4.0)
# many of those who have HSGPA = 0 have LAST_ACT_MATH_SCORE close to Max
# Cloud be a evidence their HSGPA should be NA instead of 0.

count(filter(student.record, HSGPA< 1 & LAST_ACT_MATH_SCORE >35 ))
```