Draft 1 for EDAV Project 1

Shenghan Yu

Sunday, February 07, 2016

IDEA 1: The popular tools (See Plot 1)

IDEA 2: Program Decomposition (See Plot 2)

IDEA 3: Confidence level for R user (actually we can extend that to MATLAB etc.) (See Plot 3)

IDEA 4: Compare skills, confidence level for different degree subgroups (group by PhD, IDSE, Other Masters..)

TODO: Working on using Shiny and GGPLOT2 to present the results

```
setwd("D:/Google Drive/Courses_G/1B_Exploratory Data Analysis and Visualization/edav_project1")
require(xlsx)
```

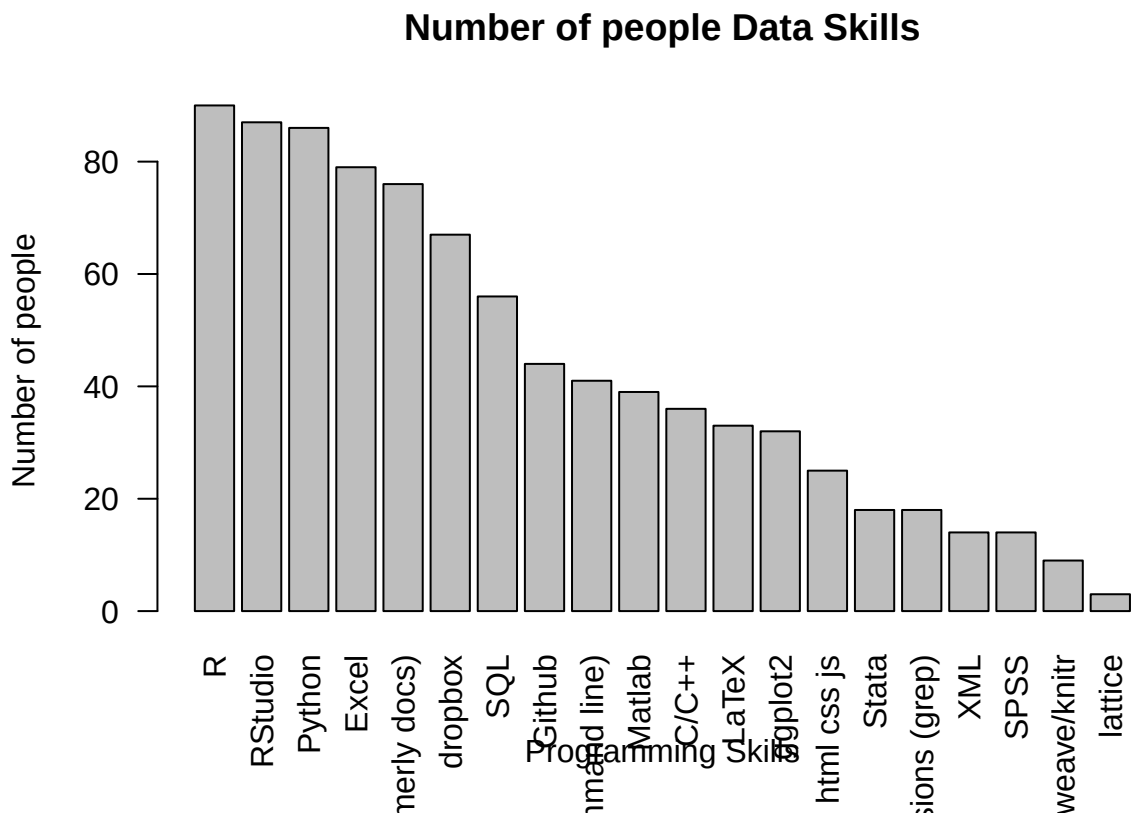
```
## Loading required package: xlsx
## Loading required package: rJava
## Loading required package: xlsxjars
```

```
library(ggplot2)
library(dummies)
```

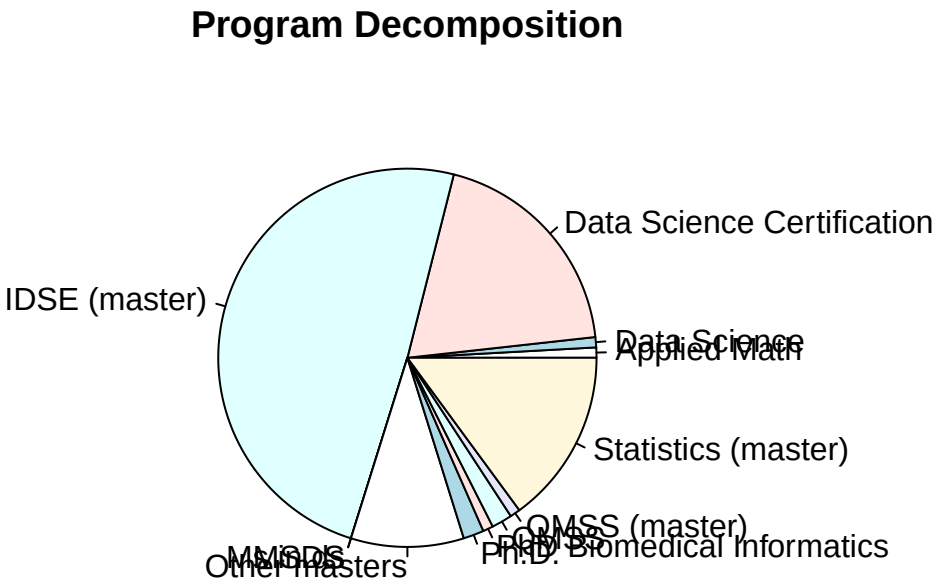
```
## dummies-1.5.6 provided by Decision Patterns
```

```
table1 <- read.xlsx("Survey+Response.xlsx", sheetName = "Form Responses 1")

# Get the skills for each samples.
skill_list <- strsplit(as.character(table1$Experiences.with.tools), ", ")

# Create a list for all skill names
skill_name <- vector(mode="character", length=0)


# Create dummies for skills, loop through samples and skills
for (sample_index in 1:nrow(table1)) {

   for (skill in skill_list[[sample_index]]) {

     # append the names, create a new column if it is a new skill
     if (!is.element(skill, skill_name)) {
       skill_name = append(skill_name,skill)
       table1[,skill] = 0
     }
```

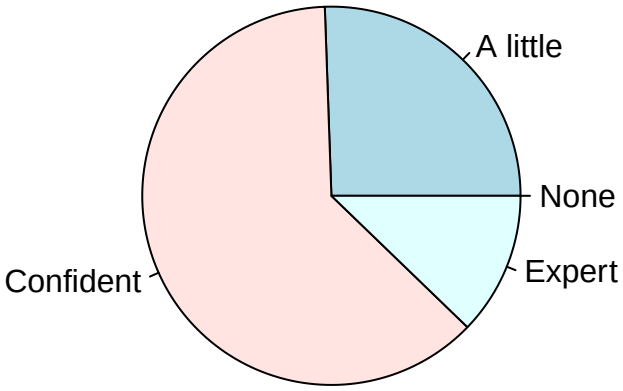
```
      table1[sample_index,skill] = 1
   }

}

# gender
table1$gender <- factor(table1$What.is.your.preferred.gender.pronoun.,levels=c("she/her","he/him","doesn
table1$gender[is.na(table1$gender)] <- 2

# plot 1: num. of people vs. skills
barplot(sort(colSums(table1[,39:58]),decreasing=T), main = 'Number of people Data Skills', las = 2, xlab
```




```
# Plot 2: program
# Clean the data
table1$Program[table1$Program == 'Ms in ds'] = 'IDSE (master)'
table1$Program[table1$Program == 'MSDS'] = 'IDSE (master)'

program <- table(table1$Program)
pie(program, main = 'Program Decomposition')
```




```r
# plot 3: Confidence
table1$conf_r <- factor(table1$Programming.and.Analytical.Experiences..R..data.manipulation.and.modelin

groupR <- data.frame(table1[table1$R==1, ])

groupR$conf_r <- factor(groupR$Programming.and.Analytical.Experiences..R..data.manipulation.and.modelin

confidencer <- table(groupR$conf_r)

pie(confidencer, main = 'Confidence level for R users')
```




```
# Correlation in R group: confident using R & number of tools
cor(as.numeric(groupR$conf_r), as.numeric(rowSums(groupR[,39:58])))
```

```
## [1] 0.4044422
```