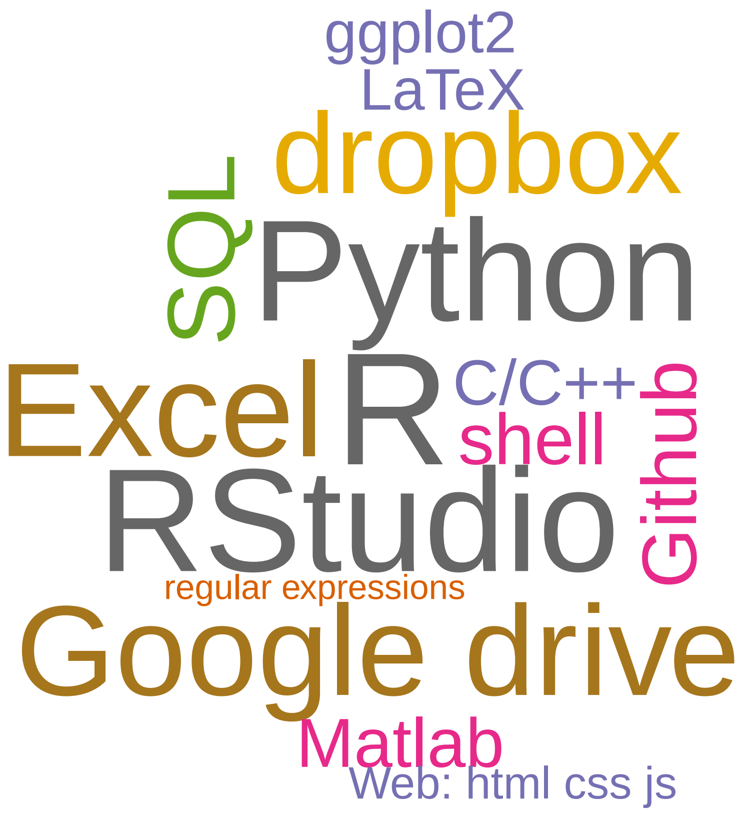
Word Cloud and Some Plots

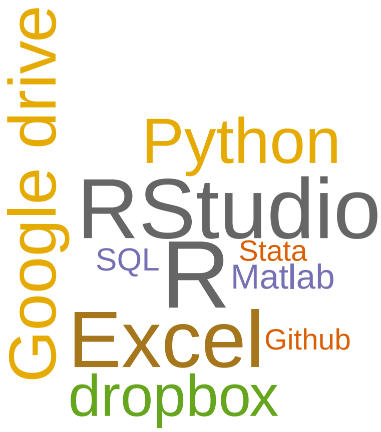
Qitong Liu

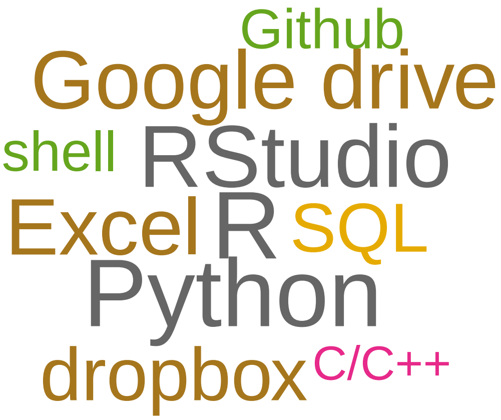
February 8, 2016

```r
library(wordcloud)
survey = read.csv("Survey+Response.csv")
words = colnames(survey)[12:31]
words[8] = "shell"
words[10] = "C/C++"
words[15] = "regular expressions"
words[16] = "Sweave/knitr"
words[18] = "Web: html css js"
words[20] = "Google drive"
freq = colSums(survey[,12:31])
pal2 <- brewer.pal(8,"Dark2")
wordcloud(words, freq, scale=c(5,.2), max.words = 15, random.order=FALSE, rot.per=.15, colors=pal2)
```



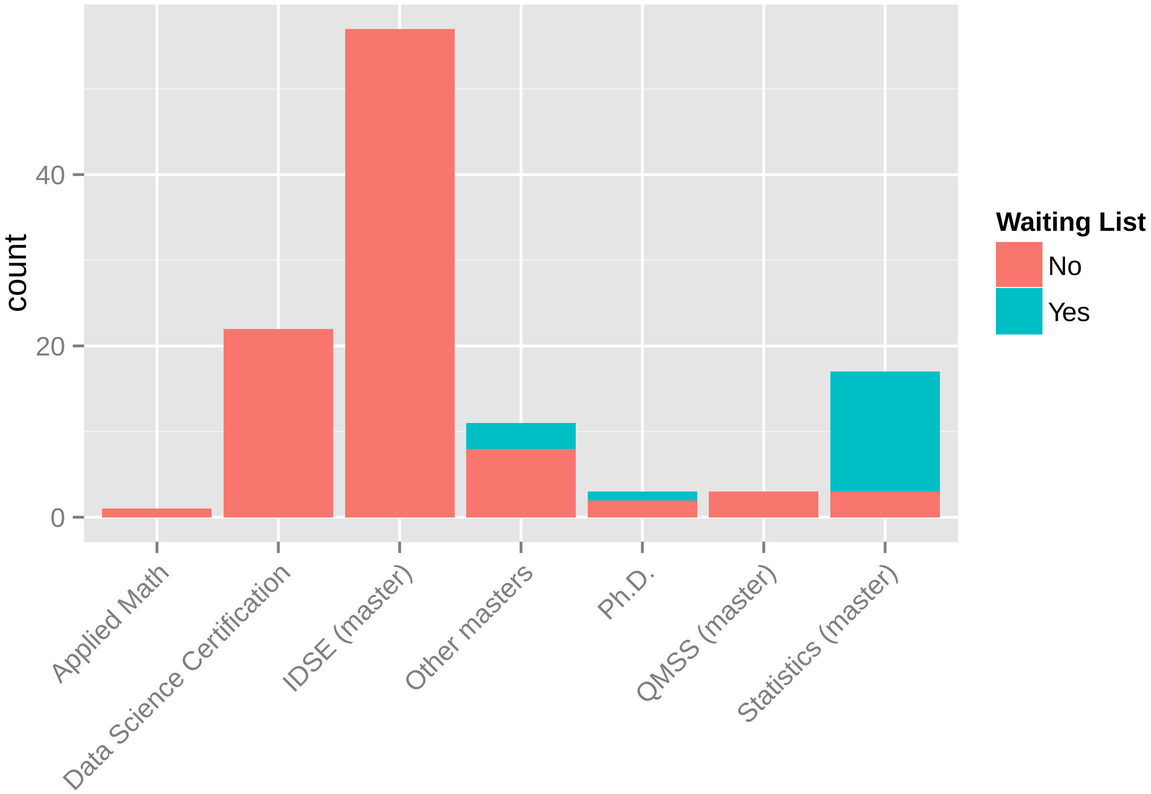
```r
words.f = survey[survey$What.is.your.preferred.gender.pronoun. == "she/her", 12:31]
words.m = survey[survey$What.is.your.preferred.gender.pronoun. == "he/him", 12:31]
freq.f = colSums(words.f)
freq.m = colSums(words.m)
par(mfrow = c(1,2))
wordcloud(words, freq.f, scale=c(3,.2), max.words = 10, random.order=FALSE, rot.per=.1, colors=pal2)
wordcloud(words, freq.m, scale=c(3,.2), max.words = 10, random.order=FALSE, rot.per=.1, colors=pal2)
```





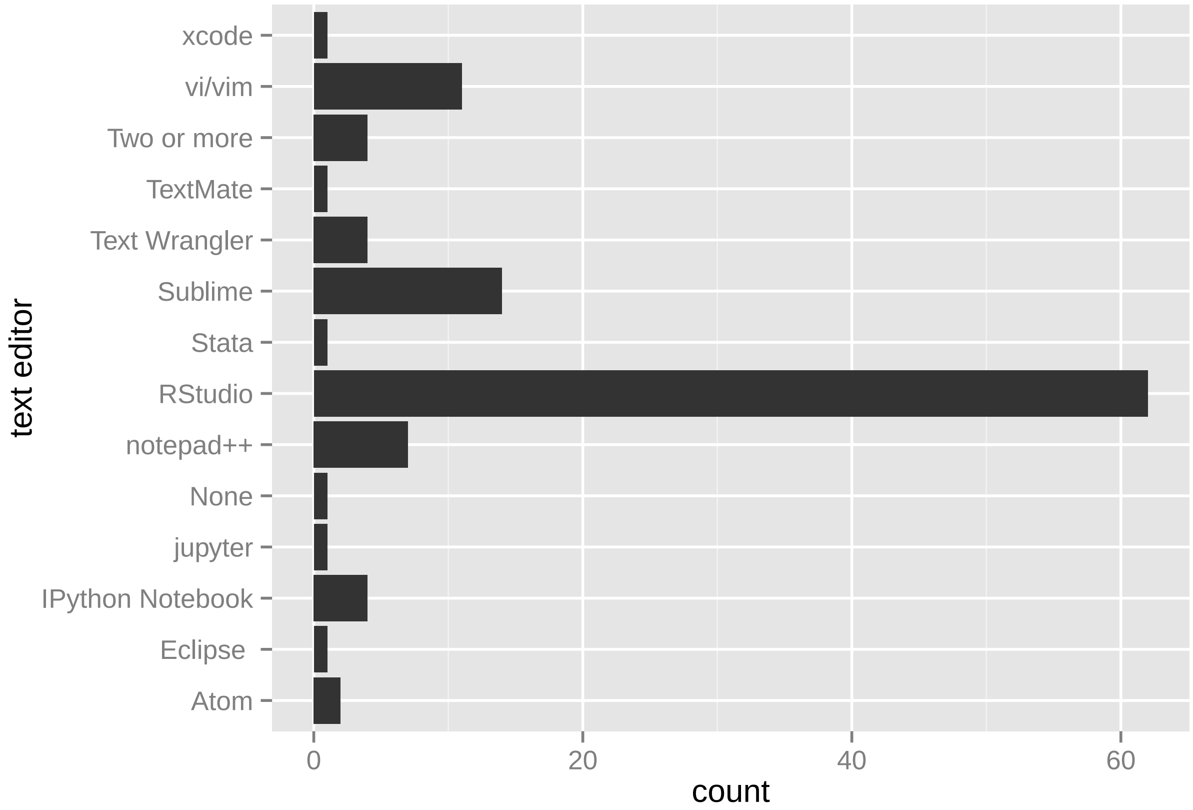
Overall, the top five popular tools are R, RStudio, Python, Excel, and Google drive. For girls and boys in this class, the top five popular tools's orders are slightly different. For girls, the order is R, RStudio, Excel, Google drive, and Python. For boys, it is R, Python, RStudio, Google drive, and Excel.

```
library(ggplot2)
# borrowed from Xuyan's code
# change factor variables into char
for(i in 1:dim(survey)[2]){
  if(class(survey[,i])=="factor"){
    survey[,i] = as.character(survey[,i])
  }
}
# clear some of the answers
# unique(survey$Program)
survey$Program[survey$Program=="MSDS"]="IDSE (master)"
survey$Program[survey$Program=="Ms in ds"]="IDSE (master)"
survey$Program[survey$Program=="Data Science"]="IDSE (master)"
survey$Program[survey$Program=="QMSS"]="QMSS (master)"
survey$Program[survey$Program=="PhD Biomedical Informatics"]="Ph.D."

# Relationship between Program and waiting list
ggplot(survey, aes(Program, fill = Are.you.on.the.waiting.list.)) + geom_histogram() + theme(axis.text.
```

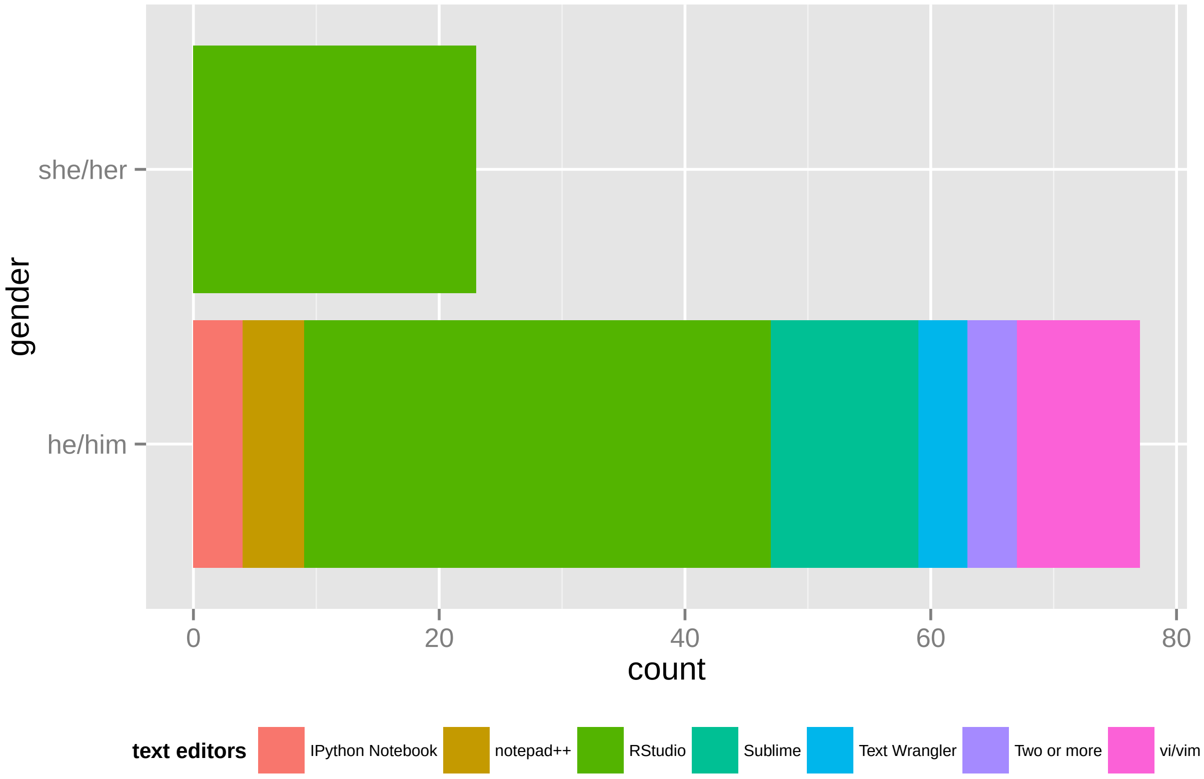


We can see that students on the waiting list are mostly Statistics(master).

```
# unique(survey$What.code.text.editor.do.you.use.most.)
survey$What.code.text.editor.do.you.use.most.[survey$What.code.text.editor.do.you.use.most. %in% c("subl
survey$What.code.text.editor.do.you.use.most.[survey$What.code.text.editor.do.you.use.most. %in% c("tex
survey$What.code.text.editor.do.you.use.most.[survey$What.code.text.editor.do.you.use.most. %in% c("ipy
survey$What.code.text.editor.do.you.use.most.[survey$What.code.text.editor.do.you.use.most. == "I used
survey$What.code.text.editor.do.you.use.most.[survey$What.code.text.editor.do.you.use.most. == "haven't
survey$What.code.text.editor.do.you.use.most.[survey$What.code.text.editor.do.you.use.most. %in% c("Ato

# Histgram of text editors
ggplot(survey, aes(What.code.text.editor.do.you.use.most.)) +geom_histogram() + coord_flip() + xlab("te
```

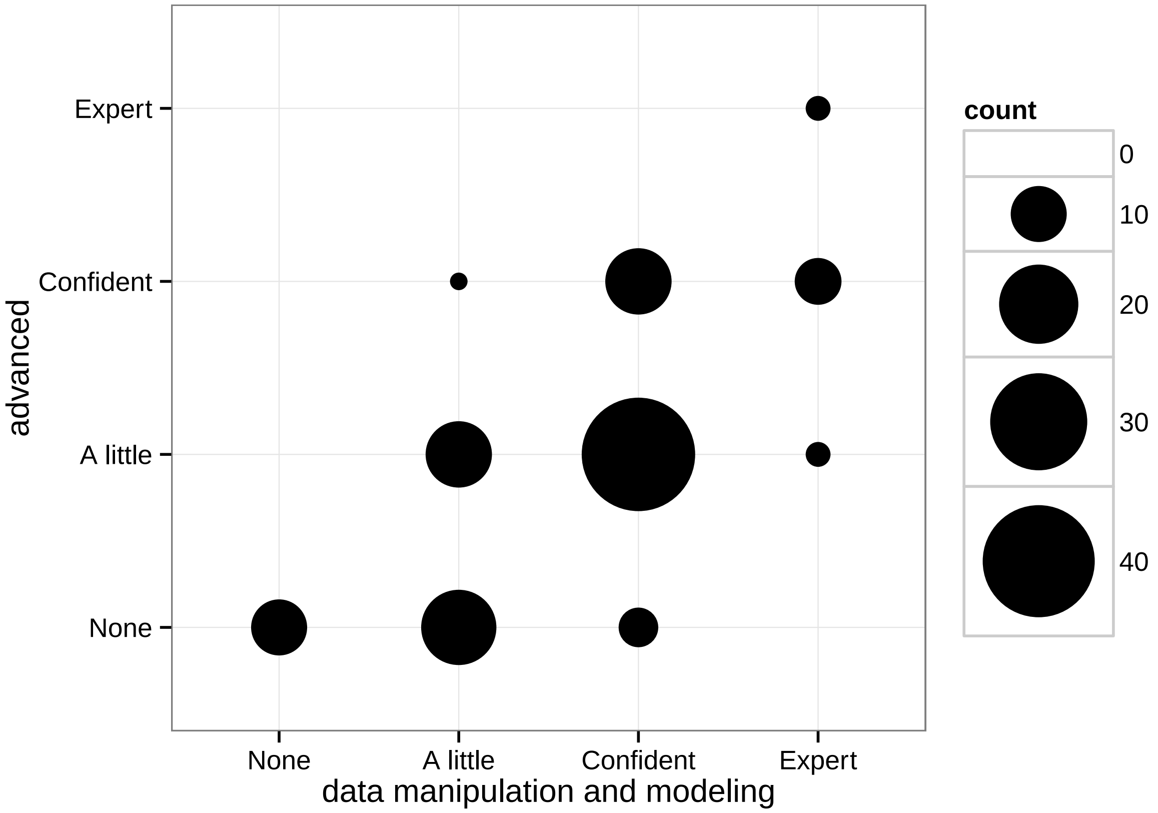


```r
editor.gender = survey[survey$What.is.your.preferred.gender.pronoun. %in% c("he/him", "she/her"), 5:6]
editor.freq = as.data.frame(table(editor.gender))

# Histgram of gender and text editors which are used by more than 2 studets
ggplot(editor.freq[editor.freq$Freq > 2,], aes(x = What.is.your.preferred.gender.pronoun., y = Freq, fil
```



Boys have a much more various choices of text editors. The only text editor that are used most by more than two girls is RStudio.

```
mani.adva = as.data.frame(table(survey[,c(4,8)]))
colnames(mani.adva) = c("data manipulation and modeling", "advanced", "count")
ggplot(mani.adva, aes(x=`data manipulation and modeling`, y=advanced, size=count)) + geom_point() + scal
```



The size of bubble represents the number of students in that category. Since using R to do data manipulation and modeling is more basic than advanced analysis, most students are confident of the former and know a little of the latter.