rpw\_final

setwd("/Users/Ryan/Desktop/Spring/EDAV/EDAV\_HW1")  
survey <- read.csv("Survey+Response.csv")  
  
#naming vars and removing empty columns  
names(survey) <- c("waitlist","program","tools","exp.Rmodeling","b5","b6","b7","b8","b9","b10","b11","gender","primaryeditor","exp.Rgraphics","exp.Radvanced","exp.documentation","exp.Matlab","exp.Github","b19","b20","b21","b22","b23","b24","b25","b26","b27","b28","b29","b30","b31","b32","b33","b34","b35","b36","b37","b38")  
survey <- survey[,c(-5:-11,-19:-38)]  
  
#dummy variables for each language/tool in tools  
tooldummies = c()  
toolList <- c("Matlab","lattice","Github","Excel","SQL","RStudio","ggplot2","shell", "C/C","Python","LaTeX","(grep)","Sweave/knitr","XML","Web: html css js","dropbox","google drive","SPSS","Stata")  
for(t in toolList){  
 tooldummies <- cbind(tooldummies,grepl(t,survey$tools))  
}  
tooldummies <- cbind(tooldummies,(grepl("R,",survey$tools)==TRUE | (grepl("R",survey$tools)==TRUE & grepl("RStudio",survey$tools)==FALSE)))  
colnames(tooldummies) <- c("Matlab","lattice","GitHub","Excel","SQL","RStudio","ggplot2","shell", "C","Python","LaTeX","grep","Sweave","XML","Web","dropbox","googledrive","SPSS","Stata","R")  
survey <- cbind(survey,tooldummies)

par(las = 2)  
par(mar=c(6,4,2,1))  
  
library(devtools)

## Warning: package 'devtools' was built under R version 3.2.3

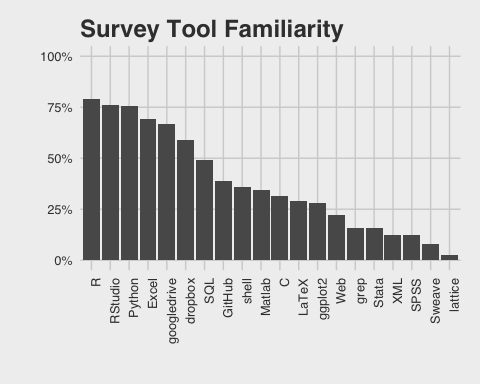
install\_github("carlosespino11/ggthemes")

## Skipping install for github remote, the SHA1 (4a6509f6) has not changed since last install.  
## Use `force = TRUE` to force installation

library(ggthemes)  
library(ggplot2)

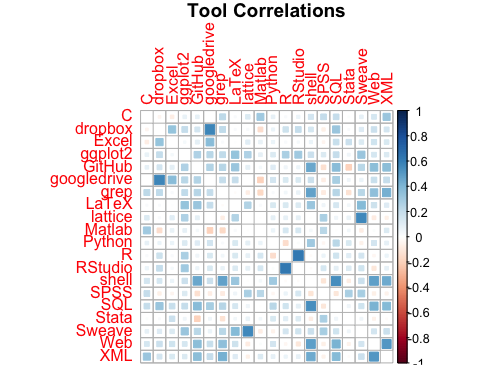
## Warning: package 'ggplot2' was built under R version 3.2.3

library(scales)  
order = sort(apply(survey[,12:31],2,mean),decreasing = TRUE)  
plotdata = as.data.frame(cbind(names(order),apply(survey[,names(order)],2,mean)))  
names(plotdata) <- c('tool','prop')  
plotdata$proportion <- as.numeric(as.character(plotdata$prop))  
plotdata$tool <- factor(plotdata$tool,levels=names(order))  
ggplot(plotdata,aes(tool,proportion)) + geom\_bar(stat = 'identity') + theme\_fivethirtyeight() + scale\_color\_tableau() + ggtitle('Survey Tool Familiarity') + xlab("") + ylab("") + scale\_y\_continuous(labels=percent, limits = c(0,1)) + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

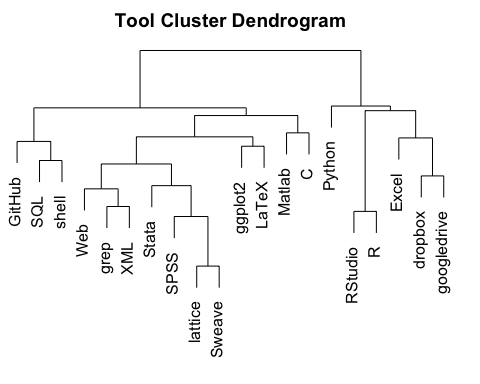


Tool Clusters

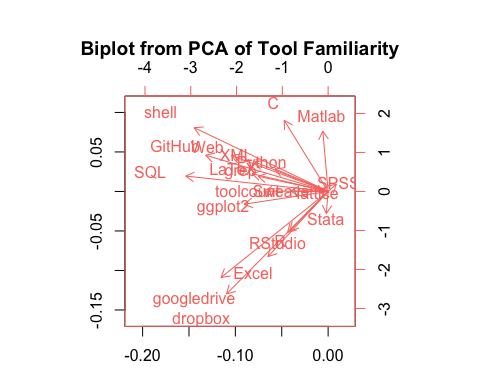
#Correlation plot  
library(corrplot)  
corrplot(cor(survey[,sort(names(order))]),method='square',diag = FALSE, mar = c(1,0,1,0),main = ' Tool Correlations')

 This table displays the pairwise correlation of familiarity with each tool. A positive correlation between a pair indicates that an individual that knows one tool means that it is more likely the individual knows the second language then the class as a whole. The table can help reveal what tools are often learned together or that may rely on each other. For example, R and RStudio has the highest (positive) correlation which is unsurprising since familiarity with RStudio would suggest familiarity with R, and the two are typically used in combination. Other highly correlated tools are google drive and dropbox (both collaboration/cloud storage services), lattice and sweave (both R libraries), and shell and SQL. Looking at whole rows (or columns) reveals which tools tend to be known by individuals with a larger set of tools. For example, SQL and shell have multiple tools with a positive correlation suggesting that if someone knows SQL or shell, they probably know a larger number of tools than the class average. Conversely, someone who knows Stata, SPSS, or MatLab probably does not know more tools than typical. Overall, most correlations are positive suggesting knowing a specific tool means that you are more likely to know a greater number of otehr tools.

#heirarchical  
dist <- dist(t(tooldummies))  
clust <- hclust(dist, method = 'average')  
par(mar = c(1,0,2,1))  
plot(clust, main = 'Tool Cluster Dendrogram', axes = FALSE)

 This dendrogram shows the agglomerative clustering using average linkage for the twenty tools. The dendrogram includes some expected pairings based on the correlation table with highly correlated pairs grouped next to each other, but it is surprising that the very first split seperates R from the different R libraries. It seems like the first split may be a split between common and uncommon tools since the six tools on the right branch are the six most commonly known tool in the survey.

#PCA biplot  
par(mar = c(3,1,4,1))  
toolcount = apply(survey[,12:31],1,function(x) mean(x))  
biplot(princomp(cbind(tooldummies,toolcount),scale = TRUE),xlabs=rep("",114), xlab = '', ylab = '',col = '#f8766d',xlim = c(-.21,.02),ylim = c(-.16,.11),main = 'Biplot from PCA of Tool Familiarity',bg = 'grey')

 This biplot is formed from a principal components analysis of the twenty tools as well as a total count of the tools. Most of the tools either point toward the top left or the bottom left, and the split is similar to the first split in the clustering with Python being a noticeable exception. The first principal component seems to be primarily overall tool expertise. This is supported by the fact that toolcount points directly to the left, and most tools point to the left half since knowing one tool means that you are more likely to know other tools. The exceptions to that are the same as from the correlation plot with Matlab, Stata, and SPSS all suggesting that familiarity with these tools doesn't make it more likely to know other tools. SQL and shell have the most negative scores on the x-axis which also agrees with the correlation plot in that knowing one of these tools suggest knowlege of a wider array of other tools. The second principal component is more difficult to interpret. Tools with high scores on the second principal component (C, shell, and Matlab) are all scripting languages that may be more common among students with an engineering background. The tools with the lowest score (googledrive, dropbox, Excel) are not scripting languages and are known by a more general group of people with or without programming experience. Then, in the middle, are statistical languages and packages such as R, Stata, and ggplot2.