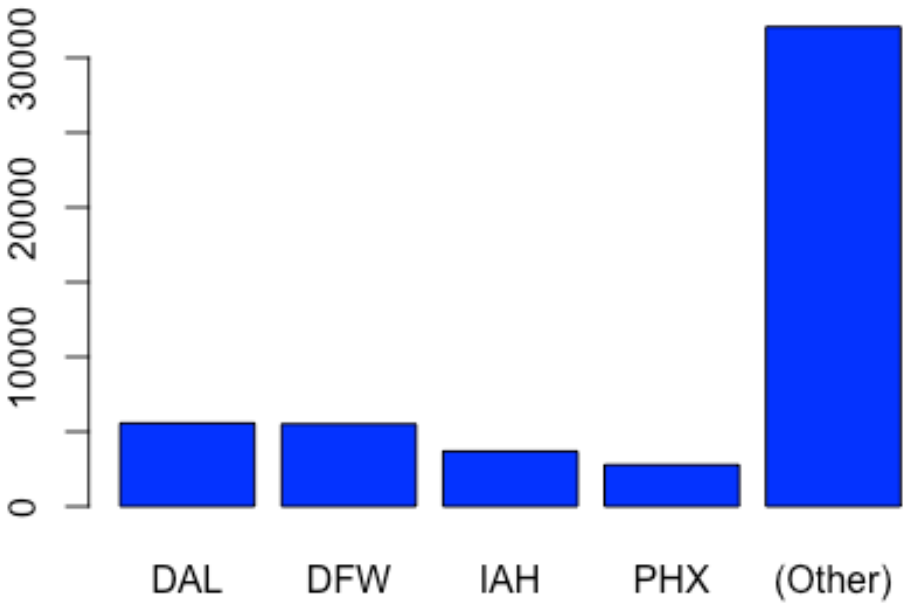
Homework 2

Caroline Nelson

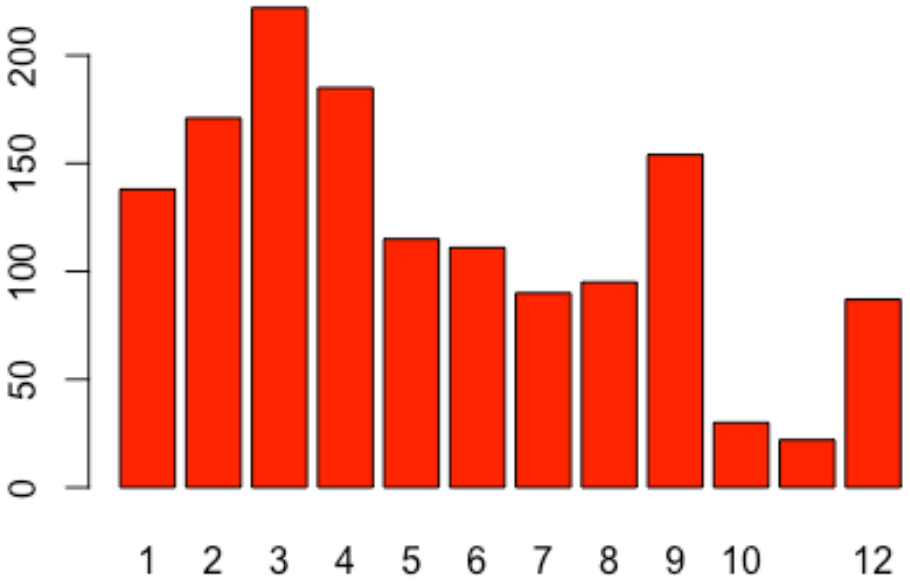
August 12, 2016

Problem 1: ABIA

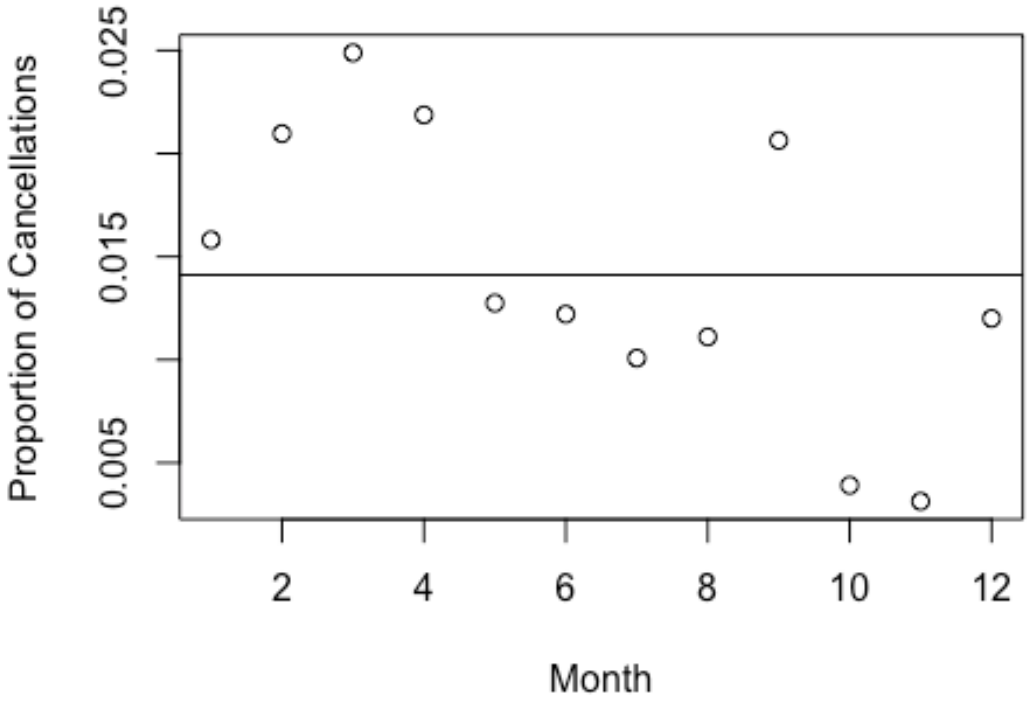
From this data set, I wanted to find out the destinations of most flights from Austin, and when the best time to travel might be, by month.



This bar plot shows the top five destinations from Austin. There are lots of 'Other' destinations, but we can see that there is a lot of business people travelling between Austin and Dallas.

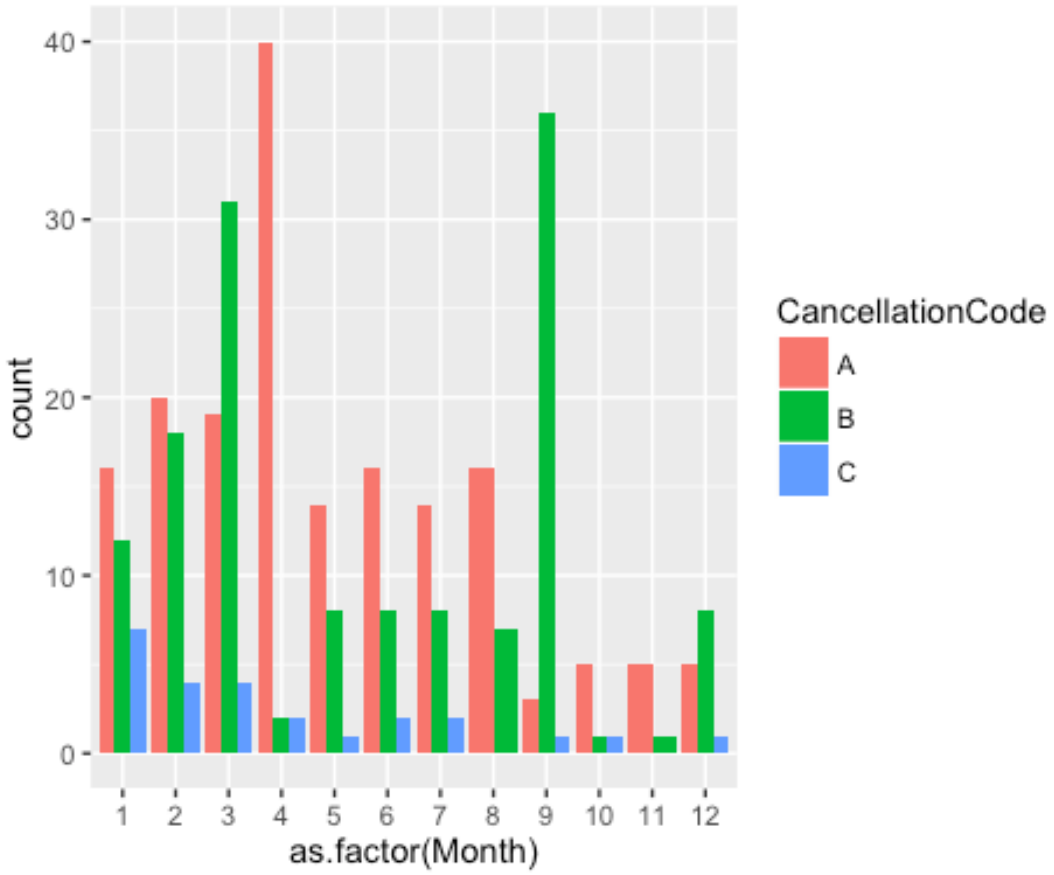


This plot shows the number of cancellations by month. We can see that there are a lot of cancellations in the month of April, but what if the majority of flights actually happened in April? To find out, I normalized each month by the number of flights in each month.



I performed a ChiSquare test to see whether one month had a particularly high amount of cancellations, and found that September, April, March, and February have significantly higher ratios of cancellations to total flights per month. So, this does mean that there are more cancellations in April than most other months. April is generally a rainy month, so let's see if rain was the reason for April's cancellations.

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```



Surprisingly, the majority of cancellations due to weather occur in September, and April has a large number of carrier cancellations. This could be due to the most common carriers being different in each month, or even destination differences.

Problem 2

Each author's documents are trained and tested with given train and test data. I created two models, Naïve Bayes and K-Nearest-Neighbors, to see which is better at predicting the author of a given topic or frequently used word.

Naive Bayes

```
library(tm)

## Loading required package: NLP

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

##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate

library(e1071)

## Warning: package 'e1071' was built under R version 3.2.2

author_dirs1 = Sys.glob('ReutersC50/C50train/*')
author_dirs2=Sys.glob('ReutersC50/C50test/*')
author_dirs = c(author_dirs1,author_dirs2)
file_list = NULL
labels = NULL
for(author in author_dirs) {
    author_name = substring(author, first=29)
    files_to_add = Sys.glob(paste0(author, '/*.txt'))
    file_list = append(file_list, files_to_add)
    labels = append(labels, rep(author_name, length(files_to_add)))
}
readerPlain = function(fname){
                readPlain(elem=list(content=readLines(fname)),
                          id=fname, language='en') }

all_docs = lapply(file_list, readerPlain)
names(all_docs) = file_list
names(all_docs) = sub('.txt', '', names(all_docs))

my_corpus = Corpus(VectorSource(all_docs))
names(my_corpus) = file_list

# Preprocessing
my_corpus = tm_map(my_corpus, content_transformer(tolower)) # make
everything lowercase
my_corpus = tm_map(my_corpus, content_transformer(removeNumbers)) #
remove numbers
my_corpus = tm_map(my_corpus, content_transformer(removePunctuation)) #
remove punctuation
my_corpus = tm_map(my_corpus, content_transformer(stripWhitespace)) ##
remove excess white-space
my_corpus = tm_map(my_corpus, content_transformer(removeWords),
stopwords("SMART"))
my_corpus <- tm_map(my_corpus, stemDocument, language = "english")
DTM1 = DocumentTermMatrix(my_corpus)
DTM1 = removeSparseTerms(DTM1, 0.975)

# Now a dense matrix
X_train = as.matrix(DTM1[1:2500,])
X_test=as.matrix(DTM1[2501:5000,])

file_names=rownames(X_train)
```

```
author_names=vector(mode='character',length=length(file_names))
for(i in 1:length(file_names)){
  temp_name=strsplit(file_names[i],'/')
  author_names[i]=temp_name[[1]][3]
}
Xtrain=data.frame(X_train,author_names)

test_names=rownames(X_test)
author_test=vector(mode='character',length=length(test_names))
for(i in 1:length(test_names)){
  temp_name=strsplit(test_names[i],'/')
  author_test[i]=temp_name[[1]][3]
}
Xtest=data.frame(X_test,author_test)

model <- naiveBayes(author_names~.,data=Xtrain)
prediction <- predict(model, newdata=Xtest)
table=table(prediction, Xtest[,ncol(Xtest)])
sum(prediction==Xtest$author_test)/length(prediction)*100
```

```
## [1] 23.48
```

Using a Naive Bayes model, we see a very low accuracy score, most likely due to assumed independence. The confusion matrix showed David Lawder, Lydia Zajc, and Tim Farrand's writings are similar to quite a few other authors' writings; they were often predicted incorrectly.

K-Nearest Neighbors

```
library(class)
set.seed(100)

# Packages
library(tm) # Text mining: Corpus and Document Term Matrix
library(class) # KNN model
library(SnowballC) # Stemming words
mat.df <- as.data.frame(data.matrix(Xtrain), stringsAsfactors = FALSE)
mat.df2 <- as.data.frame(data.matrix(Xtest), stringsAsfactors = FALSE)
# Isolate classifier
cl <- mat.df[, "author_names"]

# Create model data and remove "category"
modeldata <- mat.df[,!colnames(mat.df) %in% "author_names"]
modeldata2<-mat.df2[,!colnames(mat.df2)%in% "author_test"]
# Create model: training set, test set, training set classifier
knn.pred <- knn(modeldata, modeldata2, cl)

# Confusion matrix
conf.mat <- table("Predictions" = knn.pred, Actual = cl)
```

```r
# Accuracy
(accuracy <- sum(diag(conf.mat))/length(Xtest) * 100)
```

```
## [1] 77.44714
```

The accuracy of the KNN model was significantly better than Naive Bayes, and does well out of sample.

Problem 3

Since my initial analysis showed whole milk often on the right hand side, I narrowed my analysis down to only having whole milk on the right hand side, confidence of 50%, support=0.01, and lift greater than 2. This means that 1% of the groceries list had the left and right sides in the basket, 50% of those with the left hand side also had whole milk, and those with the left hand side are at least twice as likely to also have whole milk. I was hoping to see if people tended to buy less healthy things when buying whole milk, which is now less popular than, say, soy milk, for health enthusiasts.

```
## Warning: package 'arules' was built under R version 3.2.5

## Loading required package: Matrix

##
## Attaching package: 'arules'

## The following object is masked from 'package:tm':
##
##     inspect

## The following objects are masked from 'package:base':
##
##     abbreviate, write

## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport support minlen
maxlen
##         0.5    0.1    1 none FALSE            TRUE    0.01      1
4
##  target   ext
##   rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 98
##
```

```
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [88 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [11 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

##    lhs                                rhs          support
confidence
## 1  {curd,yogurt}                   => {whole milk} 0.01006609
0.5823529
## 2  {butter,other vegetables}       => {whole milk} 0.01148958
0.5736041
## 3  {domestic eggs,other vegetables} => {whole milk} 0.01230300
0.5525114
## 4  {whipped/sour cream,yogurt}     => {whole milk} 0.01087951
0.5245098
## 6  {other vegetables,pip fruit}    => {whole milk} 0.01352313
0.5175097
## 7  {root vegetables,tropical fruit} => {whole milk} 0.01199797
0.5700483
## 8  {tropical fruit,yogurt}         => {whole milk} 0.01514997
0.5173611
## 9  {root vegetables,yogurt}        => {whole milk} 0.01453991
0.5629921
## 10 {rolls/buns,root vegetables}    => {whole milk} 0.01270971
0.5230126
## 11 {other vegetables,yogurt}       => {whole milk} 0.02226741
0.5128806
##    lift
## 1  2.279125
## 2  2.244885
## 3  2.162336
## 4  2.052747
## 6  2.025351
## 7  2.230969
## 8  2.024770
## 9  2.203354
## 10 2.046888
## 11 2.007235
```

As it turns out, fruits and vegetables are very commonly bought with whole milk. This is most likely because these are items that tend to be bought on a weekly basis. We also see items like curd, butter, whipped/sour cream, and yogurt, which are other dairy products possibly used in similar recipes as whole milk. There does not seem to be as much of a correlation between whole milk and unhealthy products as I previously thought.