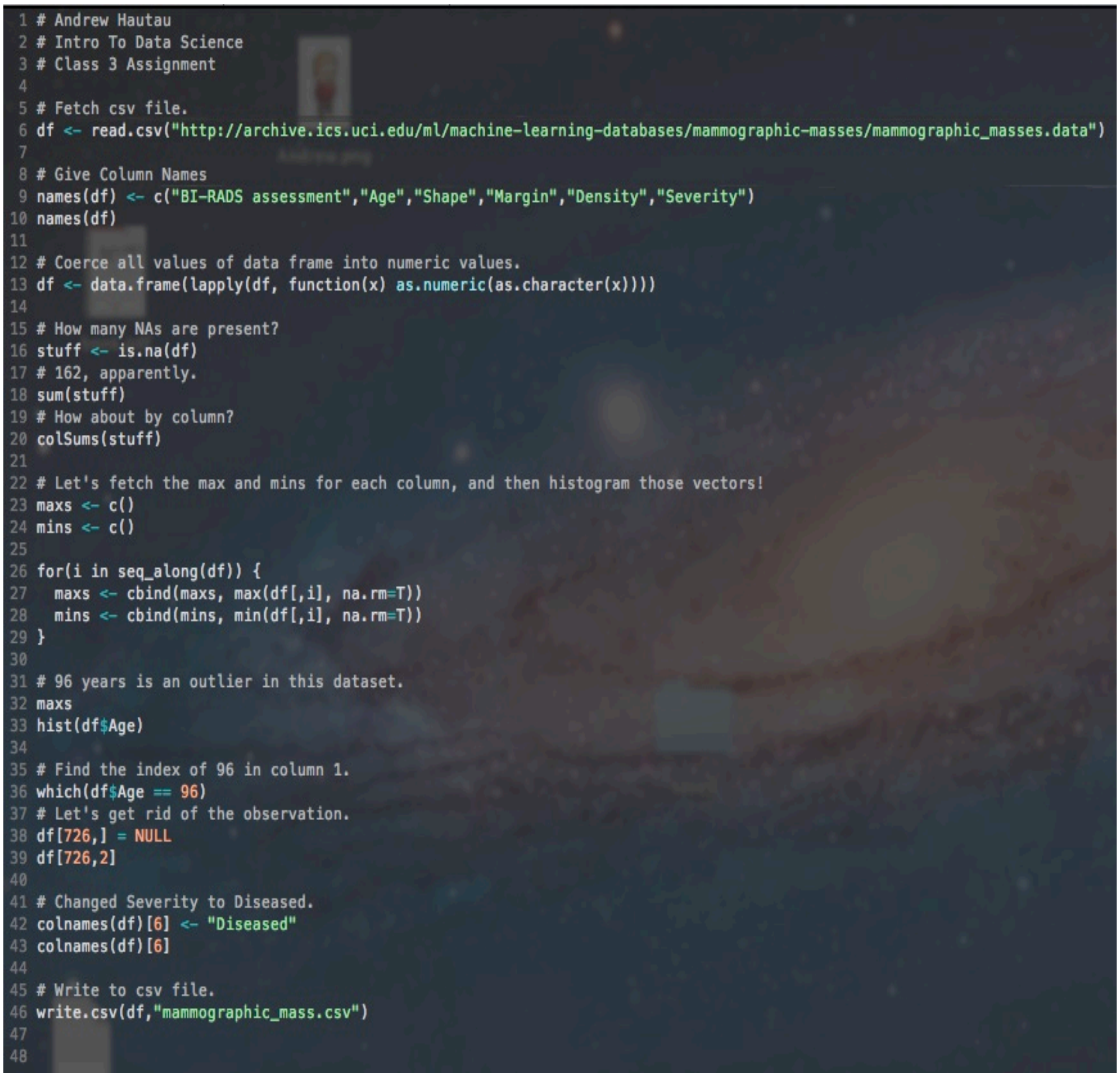
Andrew Hautau
Intro to Data Science
Week 3 Assignment

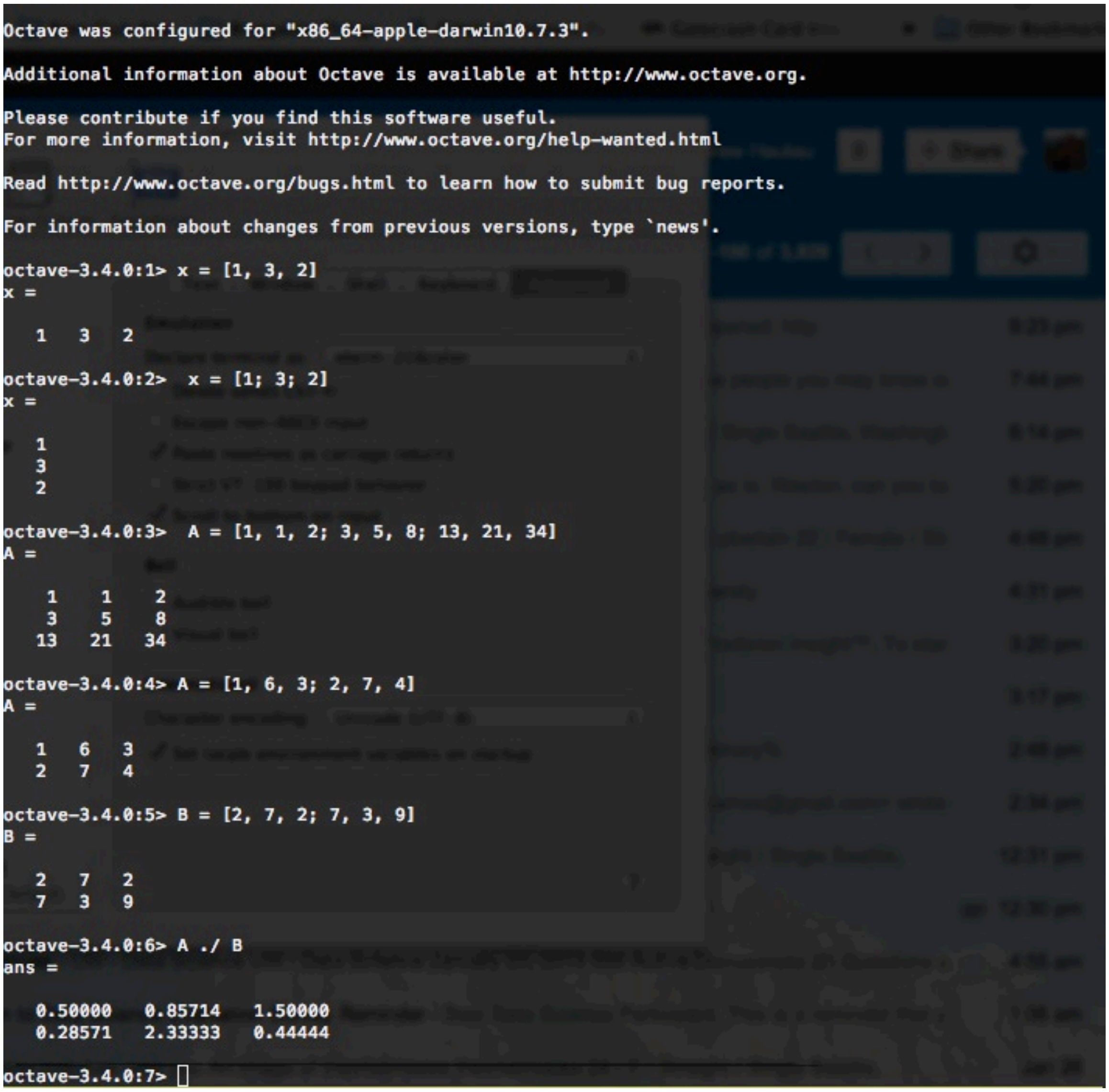
1. K-Means Clustering
   a. Data normalization is important for K-means clustering because it puts the data on equal terms.
   b. K-means methods necessarily rely on numerical data, since the method relies on finding the means of data points.
   c. Clustering is un-supervised learning because the task focuses on learning the structure of a given dataset. The task does not involve inference of a desired function from training data, as in supervised learning
2. R Code and Mammographic Masses Dataset:



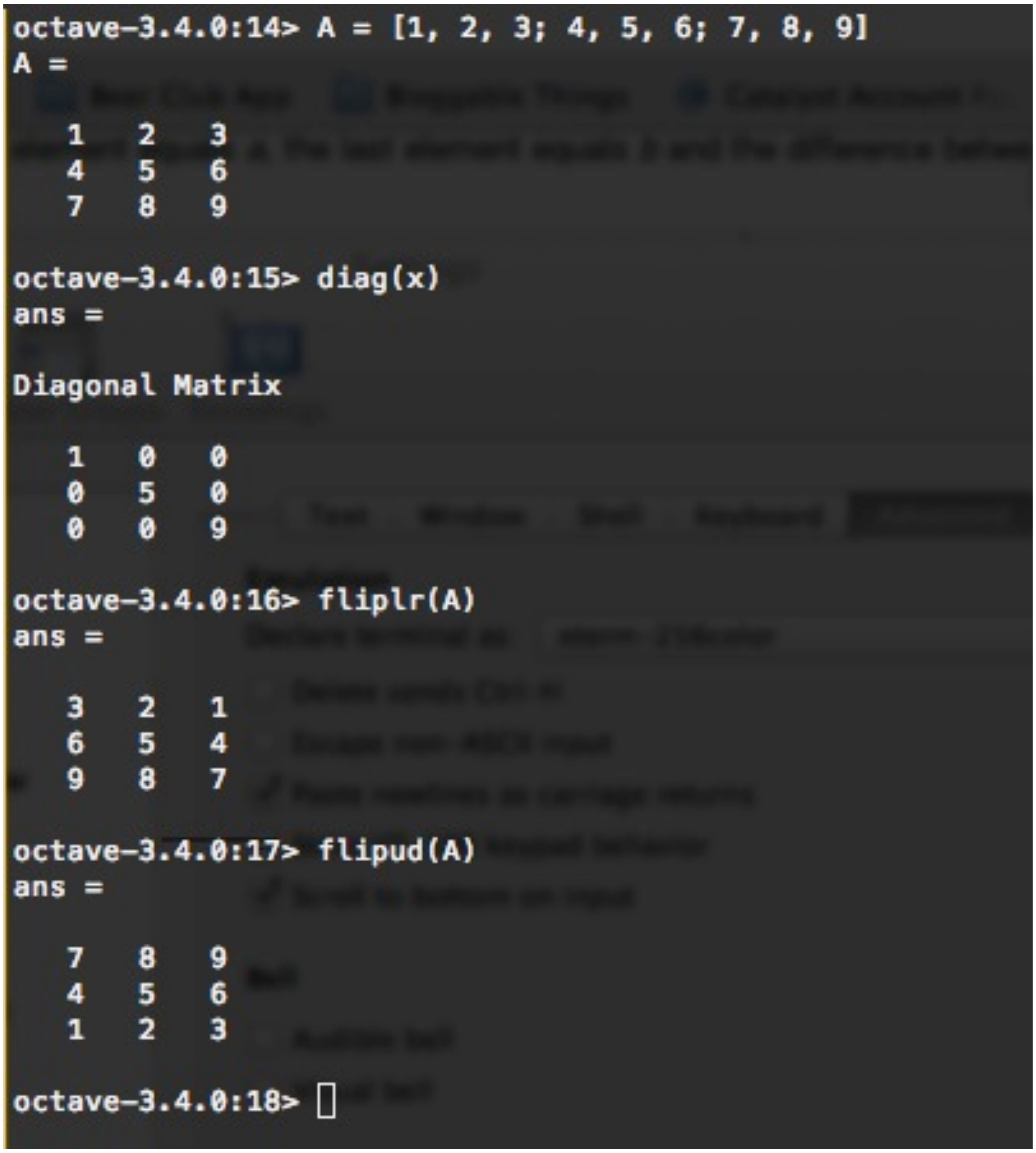
```
# Andrew Hautau
# Intro To Data Science
# Class 3 Assignment

# Fetch csv file.
df <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/mammographic-masses/mammographic_masses.data")

# Give Column Names
names(df) <- c("BI-RADS assessment","Age","Shape","Margin","Density","Severity")
names(df)

# Coerce all values of data frame into numeric values.
df <- data.frame(lapply(df, function(x) as.numeric(as.character(x))))

# How many NAs are present?
stuff <- is.na(df)
# 162, apparently.
sum(stuff)
# How about by column?
colSums(stuff)

# Let's fetch the max and mins for each column, and then histogram those vectors!
maxs <- c()
mins <- c()

for(i in seq_along(df)) {
  maxs <- cbind(maxs, max(df[,i], na.rm=T))
  mins <- cbind(mins, min(df[,i], na.rm=T))
}

# 96 years is an outlier in this dataset.
maxs
hist(df$Age)

# Find the index of 96 in column 1.
which(df$Age == 96)
# Let's get rid of the observation.
df[726,] = NULL
df[726,2]

# Changed Severity to Diseased.
colnames(df)[6] <- "Diseased"
colnames(df)[6]

# Write to csv file.
write.csv(df,"mammographic_mass.csv")


```

    a. https://github.com/CrossTheStreams/data_science/blob/master/class_3/hautau_week_3.r
    b. https://raw.github.com/CrossTheStreams/data_science/master/class_3/mammographic_mass.csv

3. Predixion Insight: After attempting into the wee-hours, I'm putting up the white flag on this one. I setup a Windows VM with Excel (also with PowerPivot), but the Predixion installation wizard would always blow up, complaining about not having an install of SQL Server. This all seems to clash with the comments on the LinkedIn group. My one theory is that SQL Server is a dependency for Insight with Excel 2010, but not other versions.
4. Playing with Octave (following tutorials from (http://en.wikibooks.org/wiki/Octave_Programming_Tutorial).



```
Octave was configured for "x86_64-apple-darwin10.7.3".

Additional information about Octave is available at http://www.octave.org.

Please contribute if you find this software useful.
For more information, visit http://www.octave.org/help-wanted.html

Read http://www.octave.org/bugs.html to learn how to submit bug reports.

For information about changes from previous versions, type `news'.

octave-3.4.0:1> x = [1, 3, 2]
x =

   1   3   2

octave-3.4.0:2>  x = [1; 3; 2]
x =

   1
   3
   2

octave-3.4.0:3>  A = [1, 1, 2; 3, 5, 8; 13, 21, 34]
A =

    1    1    2
    3    5    8
   13   21   34

octave-4.0:4> A = [1, 6, 3; 2, 7, 4]
A =

   1   6   3
   2   7   4

octave-3.4.0:5> B = [2, 7, 2; 7, 3, 9]
B =

   2   7   2
   7   3   9

octave-3.4.0:6> A ./ B
ans =

   0.50000   0.85714   1.50000
   0.28571   2.33333   0.44444

octave-3.4.0:7> 
```



```
octave-3.4.0:14> A = [1, 2, 3; 4, 5, 6; 7, 8, 9]
A =

   1   2   3
   4   5   6
   7   8   9

octave-3.4.0:15> diag(x)
ans =

Diagonal Matrix

   1   0   0
   0   5   0
   0   0   9

octave-3.4.0:16> fliplr(A)
ans =

   3   2   1
   6   5   4
   9   8   7

octave-3.4.0:17> flipud(A)
ans =

   7   8   9
   4   5   6
   1   2   3

octave-3.4.0:18> 
```