Capstone Update

Ryan Dearing

4 November 2016

R Markdown

load libraries

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(dplyr)
```

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

load each data set

```
n2005 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2005_AWARD.csv")
n2006 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2006_AWARD.csv")
n2007 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2007_AWARD.csv")
n2008 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2008_AWARD.csv")
n2009 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2009_AWARD.csv")
n2010 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2010_AWARD.csv")
n2011 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2011_AWARD.csv")
n2012 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2012_AWARD.csv")
n2013 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2013_AWARD.csv")
n2014 <- getURL("http://www.nserc-crsng.gc.ca/opendata/NSERC_GRT_FYR2014_AWARD.csv")
```

read csv and set NA values

```
n2005NA <- read.csv(text = n2005, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2006NA <- read.csv(text = n2006, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2007NA <- read.csv(text = n2007, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2008NA <- read.csv(text = n2008, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2009NA <- read.csv(text = n2009, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2010NA <- read.csv(text = n2010, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2011NA <- read.csv(text = n2011, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2012NA <- read.csv(text = n2012, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2013NA <- read.csv(text = n2013, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
n2014NA <- read.csv(text = n2014, header = TRUE, sep = ",", na.strings = c("No summary - Aucun sommaire
```

delete 2013 and 2014 extra fields to align data schema of all data sets

```
n2013NA$Num_Partie <- NULL
n2014NA$Num_Partie <- NULL
```

bind data sets

```
totalNSERC <- bind_rows(n2005NA, n2006NA, n2007NA, n2008NA, n2009NA, n2010NA, n2011NA, n2012NA, n2013NA
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

select fields of interest

```
NSERC_selected <- select(totalNSERC, Institution.Établissement, FiscalYear.Exercice.financier, AwardAmou
```

check for NAs

```
sum(is.na(NSERC_selected$Institution.Établissement) == TRUE)
```

```
## [1] 0
```

```
sum(is.na(NSERC_selected$FiscalYear.Exercice.financier) == TRUE)
```

```
## [1] 0
```

```
sum(is.na(NSERC_selected$AwardAmount) == TRUE)
```

```
## [1] 0
```

```
sum(is.na(NSERC_selected$ApplicationSummary) == TRUE)
```

```
## [1] 150592
```

there are only NA values in the "ApplicationSummary" field

filter NA values

```
NSERC_selected_filtered <- na.omit(NSERC_selected)
```

save this file locally to prevent reloading large data from the web

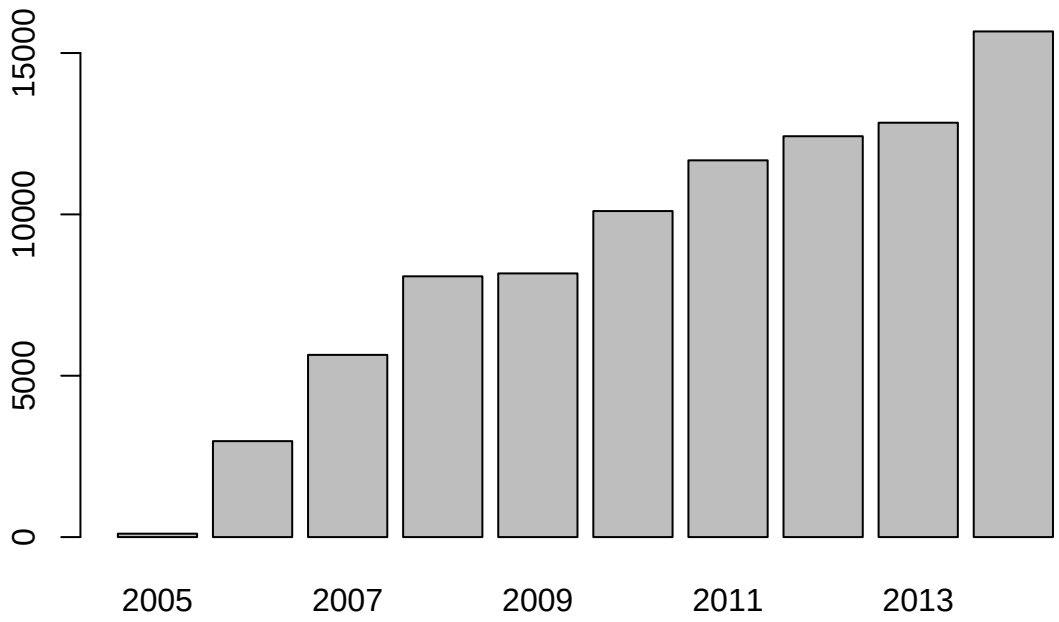
```
save(NSERC_selected_filtered, file = "NSERC.rdata")
```

load selected and filtered data from saved file

```
load("NSERC.rdata")
```

explore the data

```
barplot(table(NSERC_selected_filtered$FiscalYear.Exercice.financier)) #number of summaries are increasi
```




```
sum(as.numeric(NSERC_selected_filtered$AwardAmount)) #over 4.15 billion in funds
```

```
## [1] 4155251430
```

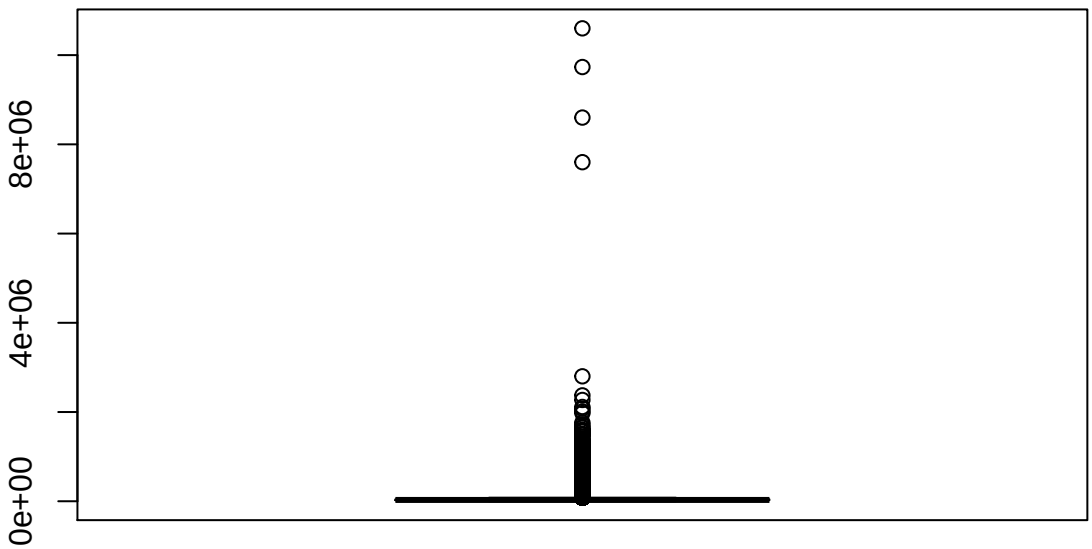
```
summary(NSERC_selected_filtered$AwardAmount)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        7    22000    29000    47390    45000 10600000
```

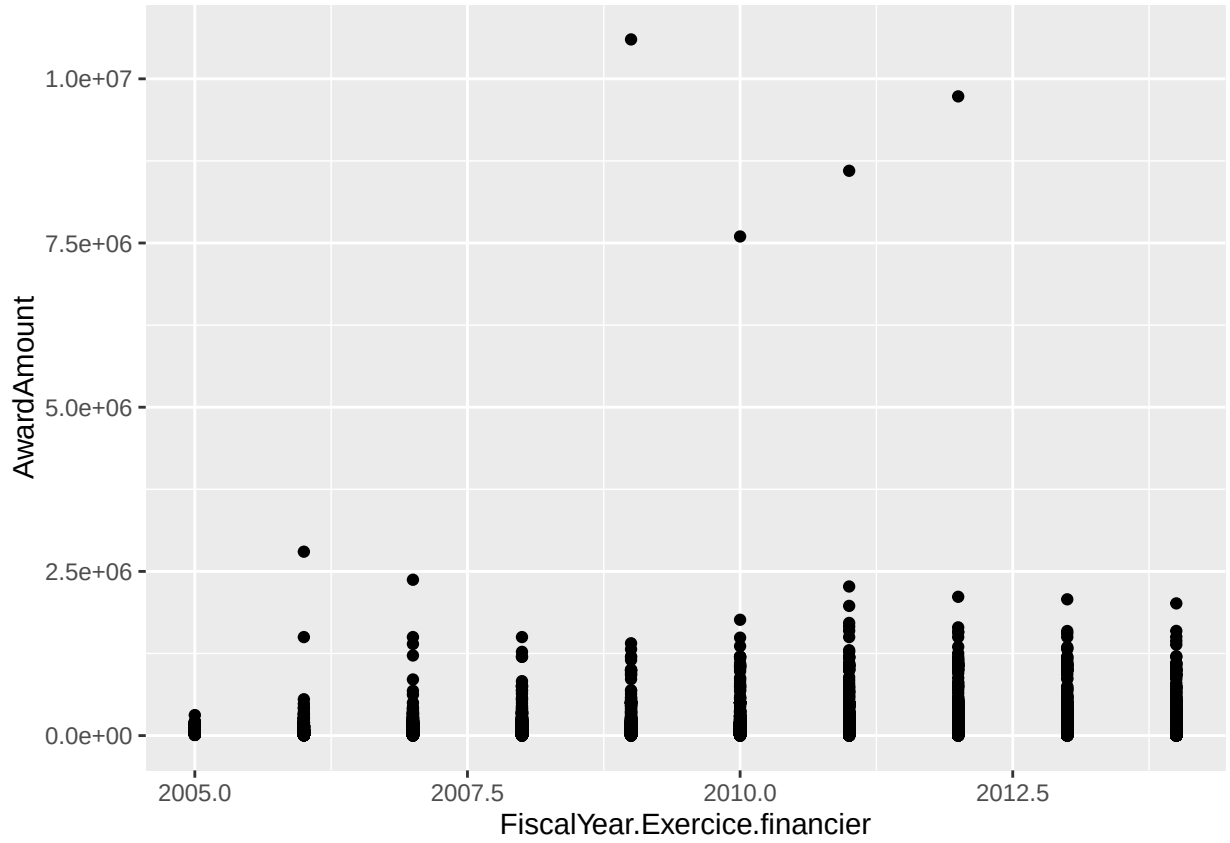
```
sd(NSERC_selected_filtered$AwardAmount) #seems too high, check for outliers
```

```
## [1] 95123.41
```

```
boxplot(NSERC_selected_filtered$AwardAmount)
```




```
ggplot(NSERC_selected_filtered, aes(x = FiscalYear.Exercice.financier, y = AwardAmount )) + geom_point()
```

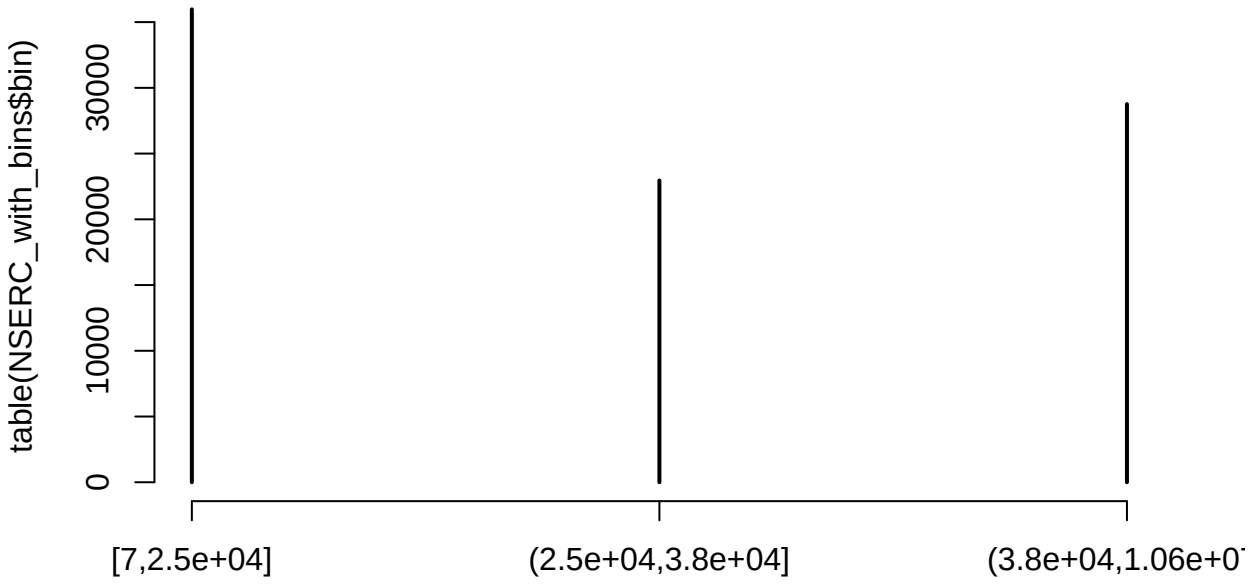


bin the data

```
q <- quantile(NSERC_selected_filtered$AwardAmount, c(0,1/3,2/3,1))
q #returns 3 nicely numbered bins, consider small, medium, and large classifiers
```

```
##        0% 33.33333% 66.66667%      100%
##         7     25000     38000  10600000
```

```
NSERC_with_bins <- data.frame(NSERC_selected_filtered, bin=cut(NSERC_selected_filtered$AwardAmount, q, i
plot(table(NSERC_with_bins$bin))
```


table(NSERC_with_bins$bin)
0
10000
20000
30000
[7,2.5e+04]
(2.5e+04,3.8e+04]
(3.8e+04,1.06e+0

Summary Findings:

I cleaned my data and trimmed it down to 4 attributes, 2 of primary interest ApplicationSummary (text) and AwardAmount (integer) and 2 attributes that may prove interesting in post analysis. I then binned the data, adding a fifth attribute. After filtering the data down, there are 87,684 records that include an ApplicationSummary out of 238,276 initial records. An interesting trend in the data is that the number of applications that include a summary have been increasing each year. The total amount of funds distributed (to applications that included a summary) is in excess of $4.15 billion. AwardAmounts (with summaries) range from $7 up to $10.6 million. The mean award amount is $47,390 with a standard deviation of $97,123. The large standard deviation and low median ($29,000) are due to 4 extreme outliers. These 4 cases are each over $7.5 million, while every other data point is below $3 million. Based on these statistics I am defining Small, Medium, and Large awards as; less than $25,000, between $25,000 and $38,000, and greater than $38,000 respectively.

Questions and Guidance:

The outliers are pretty extreme and are distorting the data quite a bit. Is this a concern with the type of analysis we are doing? Should I exclude them and use new summary statistics to recalculate my bins? In your opinion, have I chose my bins properly? Should I add a 4th bin of Very Large for grants of over $1 million?

Some of the summaries are in French. Does this matter for SVM and Naive Bayes? Are there NLP algorithms that work on French? Should I include the French records or should I simply exclude them?

Any other insights and feedback are welcome. Also, please let me know if I have missed anything for this phase or if you need anything more.