MonarchExplore

Jerad Acosta

March 6, 2015

Libraries

```
library(ggplot2)
require(caret)
require(dplyr)
require(reshape2)
```

Import Data

```
# Set working Directory
wd <- "/Users/irJERAD/1DataContest/Monarch"
setwd(wd)

# load data into R for Manipulation
filePath <- "/Users/irJERAD/1DataContest/Monarch/MonarchPunishmentsEditCSV.csv"
data <- read.csv(filePath)

# Create Directory for data as we munge
if(!file.exists("./RData")){dir.create("./RData")}
```

Summarize Data

```
str(data)
```

```
## 'data.frame':    984 obs. of  18 variables:
##  $ id                                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ timestamp                           : Factor w/ 61 levels "","00:00.0","01:00.0",..: 58 4 54 46 29
##  $ student_id                          : int  9 80 80 80 80 80 80 80 80 80 ...
##  $ grade                               : int  2 4 4 4 4 4 4 4 4 4 ...
##  $ date_and_time_of_misbehavior        : Factor w/ 608 levels "","00:00.0","01:00.0",..: 129 262 223
##  $ location_of_misbehavior             : Factor w/ 131 levels "","12th and Imperial",..: 26 93 93 10
##  $ documenting_staff_id                : int  1 2 2 2 2 2 2 2 2 2 ...
##  $ documenting_staff                   : Factor w/ 13 levels "","Mr. Abdi",..: 2 4 4 4 4 4 4 4 4 4 .
##  $ classroom_or_administrative_managed : Factor w/ 3 levels "","Administrative",..: 2 2 2 2 2 2 2 2 1
##  $ type_of_misbehavior                 : Factor w/ 249 levels "","Attendance (CM), Verbal/Physical In
##  $ narrative_description_of_misbehavior: Factor w/ 851 levels "","\"Borrowed\" another students phone
##  $ reporting_staff_id                  : int  157 45 7 26 91 32 112 112 112 112 ...
##  $ reporting_staff                     : Factor w/ 67 levels "","FIT","Front Desk",..: 58 10 5 7 38 1
##  $ d12_planning_completed              : Factor w/ 7 levels "","A, B, C","A, B, C, D",..: 3 3 3 3 3 3
##  $ narrative_of_consequence            : Factor w/ 453 levels "","1/2 day in d12",..: 395 1 38 1 1 1
##  $ consequence                         : Factor w/ 89 levels "","After School Detention",..: 14 56 56
##  $ Consequences_transcribed            : Factor w/ 65 levels "","C","C, LD",..: 11 43 43 43 43 43 43
##  $ misbehavior_transcribed             : Factor w/ 170 levels "","C","C , MI , NC , DS , I",..: 24 14
```

** FOR TEMP REFERENCE, REMOVE WHEN DONE** 'data.frame': 984 obs. of 18 variables:
$ id : int 1 2 3 4 5 6 7 8 9 10 . . .

$ timestamp : Factor w/ 61 levels “”,“00:00.0”,“01:00.0”,..: 58 4 54 46 29 40 61 18 38 31 . . .
$ student_id : int 9 80 80 80 80 80 80 80 80 80 . . .
$ grade : int 2 4 4 4 4 4 4 4 4 4 . . .
$ date_and_time_of_misbehavior : Factor w/ 608 levels “”,“00:00.0”,“01:00.0”,..: 129 262 223 425 366 373 458 472 481 491 . . .
$ location_of_misbehavior : Factor w/ 131 levels “”,“12th and Imperial”,..: 26 93 93 10 123 97 10 10 10 10 . . .
$ documenting_staff_id : int 1 2 2 2 2 2 2 2 2 2 . . .
$ documenting_staff : Factor w/ 13 levels “”,“Mr. Abdi”,..: 2 4 4 4 4 4 4 4 4 4 . . .
$ classroom_or_administrative_managed : Factor w/ 3 levels “”,“Administrative”,..: 2 2 2 2 2 2 2 2 2 2 . . .
$ type_of_misbehavior : Factor w/ 249 levels “”,“Attendance (CM), Verbal/Physical Intimidation (D12)”,..: 32 192 119 89 103 152 90 241 119 89 . . .
$ narrative_description_of_misbehavior: Factor w/ 851 levels “”,“"Borrowed" another students phone and showed a personal video. The other student got very upset and it created a scene in th”| **truncated**,..: 743 525 618 712 780 533 154 589 98 72 . . .
$ reporting_staff_id : int 157 45 7 26 91 32 112 112 112 112 . . .
$ reporting_staff : Factor w/ 67 levels “”,“FIT”,“Front Desk”,..: 58 10 5 7 38 11 52 52 52 52 . . .
$ d12_planning_completed : Factor w/ 7 levels “”,“A, B, C”,“A, B, C, D”,..: 3 3 3 3 3 3 3 3 3 3 . . .
$ narrative_of_consequence : Factor w/ 453 levels “”,“1/2 day in d12”,..: 395 1 38 1 1 1 1 1 1 1 . . .
$ consequence : Factor w/ 89 levels “”,“After School Detention”,..: 14 56 56 56 56 56 56 56 56 56 . . .
$ Consequences_transcribed : Factor w/ 65 levels “”,“C”,“C, LD”,..: 11 43 43 43 43 43 43 43 43 43 . . .
$ misbehavior_transcribed : Factor w/ 170 levels “”,“C”,“C , MI , NC , DS , I”,..: 24 149 86 59 75 109 60 20 86 59 . . .

- 61 levels of timestamps and only 60 hours in a day.
  + Find and remove factor level “”

- Grade level is an int variable
  + cast as factor

```
# find and remove empty factor level
r <- which(data$timestamp == "")
data[r,]
```

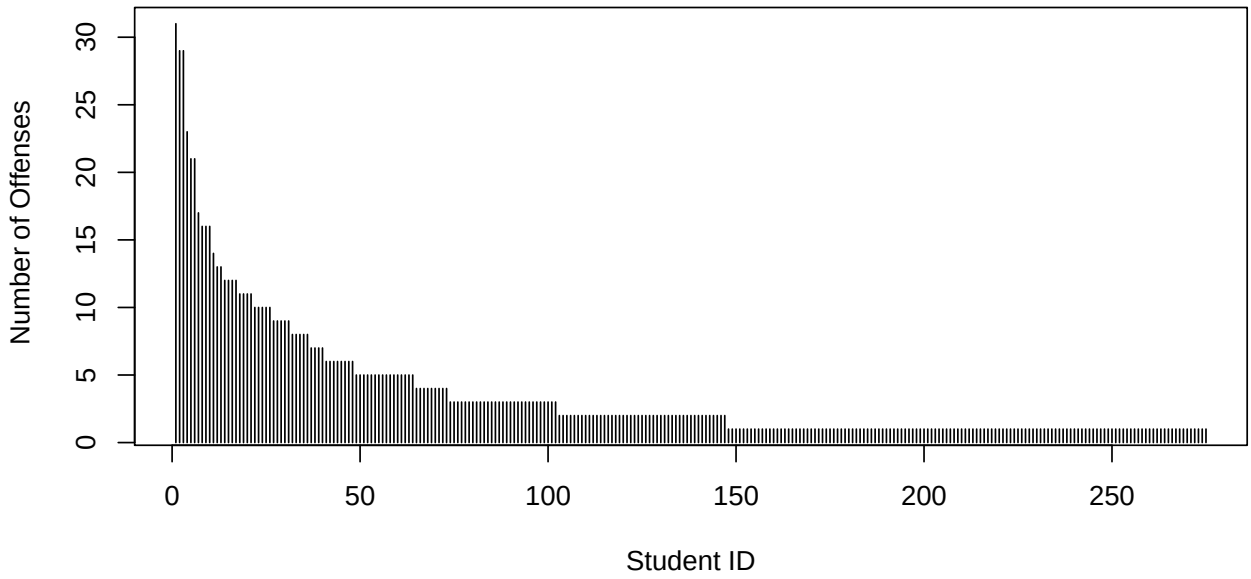
```
##     id timestamp student_id grade date_and_time_of_misbehavior
## 984 NA                   NA    NA
##     location_of_misbehavior documenting_staff_id documenting_staff
## 984                                          NA
##     classroom_or_administrative_managed type_of_misbehavior
## 984
##     narrative_description_of_misbehavior reporting_staff_id
## 984                                                    NA
##     reporting_staff d12_planning_completed narrative_of_consequence
## 984
##     consequence Consequences_transcribed misbehavior_transcribed
## 984
```

row 984 is an empty row.
Remove it from the dataset

```
data <- data[-c(r),]
```

Explore density of distribution amongst students

```
# Number of events per student table
eventTable <- table(data$student_id)
# Create a histogram counting number of behavioral events per student
plot(eventTable[order(-eventTable)], type = 'h', xlab = "Student ID", ylab = " Number of Offenses")
```




```
quantile(eventTable)
```

```
##   0%  25%  50%  75% 100%
##    1    1    2    4   31
```

```
# percentage with 4 or fewer offenses
fourPcnt <- (sum(eventTable <= 4) / length(eventTable)) * 100
fourPcnt
```

```
## [1] 76.72727
```

Here we see the majority of behavioral offenders are have 2 or fewer offenses.
76.73% of reported offenders have 4 or fewer.
As it turns out, 48 of the 275 students account for over 50% of the offenses recorded.
In other words 17.45% of the offending students are responsibly for 57.17% of the offenses.

Using this distribution data we can examine 2 population of offenders to search for clues about what separates or lays at the center of these populations.

Looking at 2 populations of Offenders

```
sum(eventTable <= 3) / length(eventTable)
```

```
## [1] 0.7345455
```

```
# number of Students with over 10 offenses
sum(eventTable > 10)
```

```
## [1] 21
```

The Population of students who have 3 or fewer offenses and make up 73.45% of the offending students

The population of students with more than 10 offenses makes up for over one-third of all the offenses recorded. To be precise they make up 35.71% of the behavioral offenses.

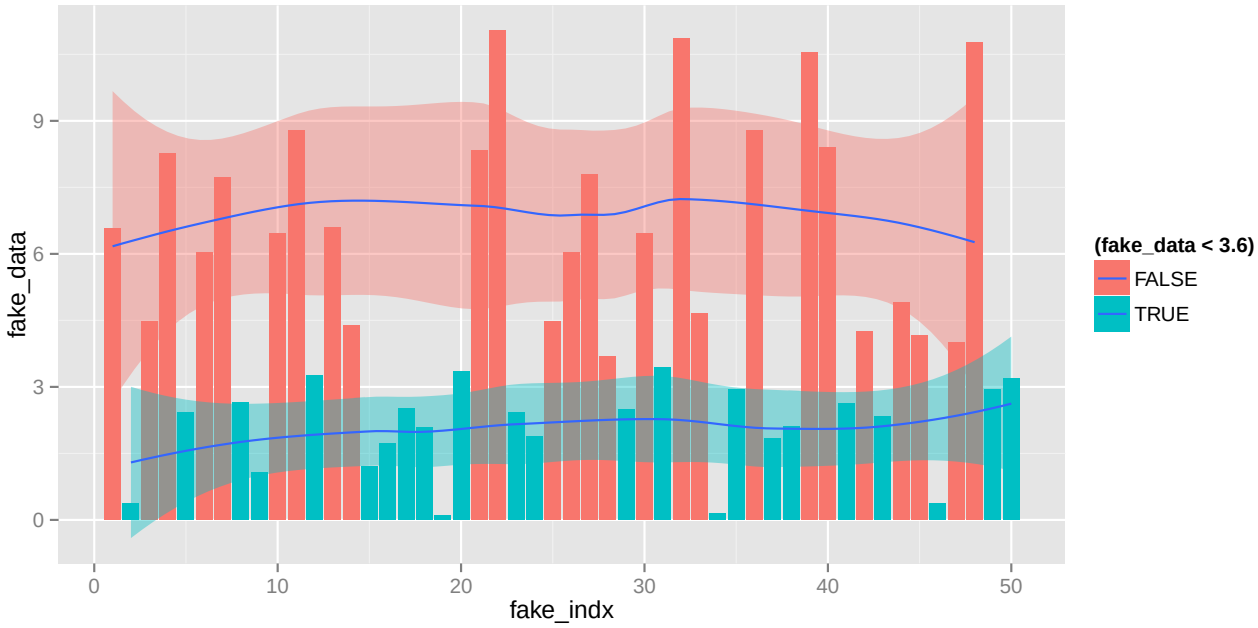
```
round(sum(head(eventTable[order(-eventTable)], n = sum(eventTable > 6))) / sum(eventTable) * 100 ,2)
```

```
## [1] 52.29
```

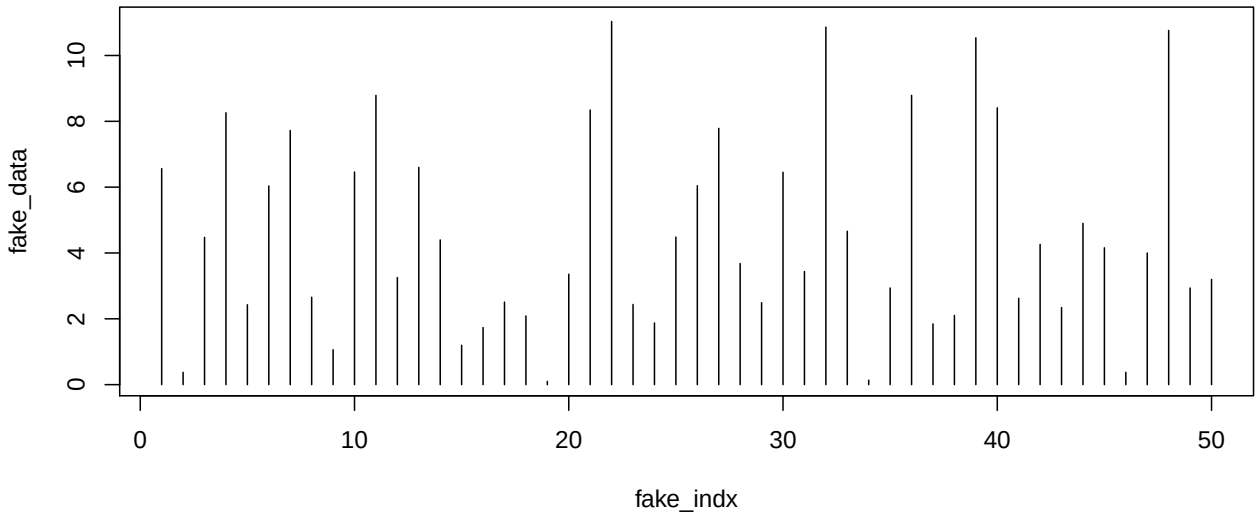
Students with more than 6 offences make up apx 52.29% of the population of offenses.

```
fake_data <- abs(rnorm(50, mean = 3.6, sd =  4.7))
fake_indx <- c(1:50)
fake_frame <- data.frame(fake_indx, fake_data)
fake_plot <- ggplot(fake_frame, aes(fake_indx, fake_data, fill = (fake_data < 3.6)))
fake_plot + geom_bar(stat = "identity") + geom_smooth()
```

```
## geom_smooth: method="auto" and size of largest group is <1000, so using loess. Use 'method = x' to cl
```




```
plot(fake_frame, type = "h")
```



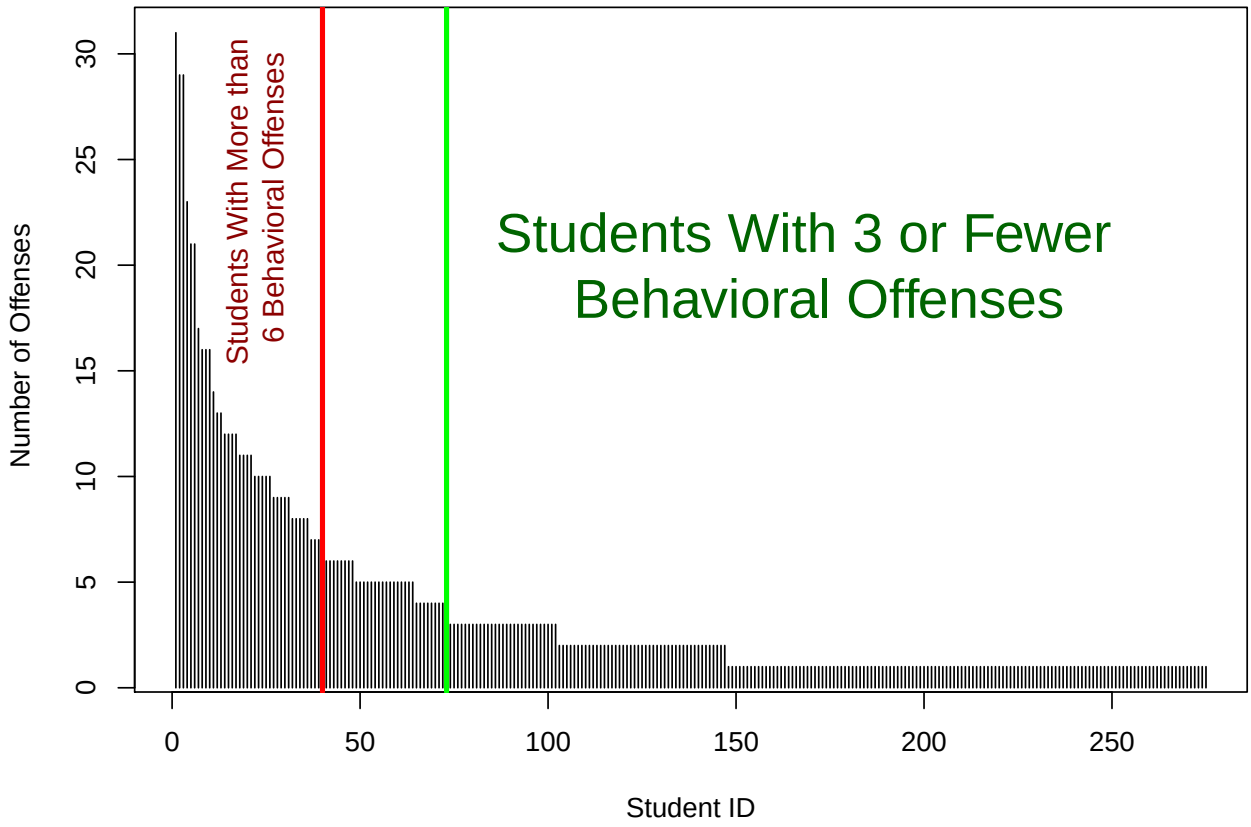
Index Populations

```
# Index of Student with more than 10 offenses
overTenIndx <- names(eventTable[eventTable > 10])
tenData <- filter(data, student_id %in% overTenIndx)

overSixIndx <- names(eventTable[eventTable > 6])
sixData <- filter(data, student_id %in% overSixIndx)

# Index of Students with 3 or less offenses
threeLessIndx <- names(eventTable[eventTable <= 3])
threeData <- filter(data, student_id %in% threeLessIndx)

# review plot with line at 6 or more offences
plot(eventTable[order(-eventTable)], type = 'h', xlab = "Student ID", ylab = " Number of Offenses")
# add line differentiation population with more than 6 offenses
abline(v = length(eventTable) - sum(eventTable <= 3),
       col = "green", lwd = 3,
       text(170,20, labels = "Students With 3 or Fewer \n Behavioral Offenses",
            col = "Dark Green", cex = 2))
# add line differentiating population with 3 or less offenses
abline(v = sum(eventTable > 6),
       col = "red", lwd = 3,
       text(22,23, labels = "Students With More than\n 6 Behavioral Offenses",
            col = "Dark red", srt = 90, cex = 1.1))
```




```
# Create Data frame from organized table
table_frame <- as.data.frame(eventTable[order(-eventTable)])
```

Transcibing Variables

We want to make information in our data easier for exploration and statistical analysis. To do so, we will create some factor variables and condense certain variable observations from things like "4th grade class" and "3rd grade Classroom" to just "Class" or "Classroom".
This makes sense because from the student's perspective both of these nominal categories would be associated with the environment of a classroom.

```
# Create environment factor variable
# Search for word class to create a classroom level in the environmental factors
# Note for personal learning: Which ever factor level is assigned first is the one that stays
# Just like an If / Else statement - Priority goes to first statement
data$environment <- ifelse(grepl("class", data$location_of_misbehavior, ignore.case = TRUE), "Classroom
                           ifelse(grepl("PE", data$location_of_misbehavior, ignore.case = TRUE), "PE",
                                  ifelse(grepl("lunch", data$location_of_misbehavior, ignore.case = TRU
                                         ifelse(grepl("yoga", data$location_of_misbehavior, ignore.case
                                                ifelse(grepl("gym", data$location_of_misbehavior, ignor
                                                       ifelse(grepl("blacktop", data$location_of_misbeha
                                                              ifelse(grepl("cafeteria", data$location_o
                                                                     ifelse(grepl("hallway", data$locati
                                                                            ifelse(grepl("playground",
                                                                                   ifelse(grepl("snack"
```