Project 1

Aaron Niskin, Charlie Edelson, Chris Leonard

October 26, 2016

```r
library(data.table)
setwd("~/Documents/courses/newCollege/current/eda/projects/project1")
```

```r
r_formatting_sucks <- function(n) {
  if(n < 10) {
    return(paste("0", n, sep=""))
  }
  else {
    return(n)
  }
}
getFileName <- function(startYear) {
  R_is_stupid <- paste("ignore/MERGED", startYear, "_", r_formatting_sucks((startYear + 1) %% 100), "_P
  return(R_is_stupid)
}
getFilesColumns <- function(years, cols, nullVals) {
  yearlyDats <- lapply(years, function(year){
    tmp <- read.csv(getFileName(year), na=nullVals, stringsAsFactors = FALSE)
    tmp <- tmp[,cols]
    tmp$startYear = year
    return(tmp)
  })
  tmp <- yearlyDats[1]
  lapply(yearlyDats[2:length(yearlyDats)], function(dats) {
    tmp <- rbind(tmp, dats)
  })
  return(tmp)
}
```

```r
dataStartYears <- 1996:2014
csvNullVals = c("NULL", "PrivacySuppressed")
famDf = getFilesColumns(dataStartYears, c("FAMINC", "MD_EARN_WNE_P10"), csvNullVals)
#df = read.csv(getFileName(dataStartYears[1]), na=c("NULL", "PrivacySuppressed"), stringsAsFactors = FA
#df$startYear = dataStartYears[1]
#for (year in dataStartYears[dataStartYears > dataStartYears[1]]){
#  tmp <- read.csv(getFileName(year), na=c("NULL", #"PrivacySuppressed"), stringsAsFactors = FALSE)
#  tmp$startYear = year
#  df <- rbind(df, tmp)
#}
#write.csv(df, "ignore/complete_data.csv", na="")
#rm(tmp)
```

```r
df = read.csv(getFileName(2005), na=c("NULL", "PrivacySuppressed"), stringsAsFactors = FALSE)
unique(df$C200_4)
```

```
df <- as.data.table(read.csv("ignore/complete_data.csv"))
```

```
df <- as.data.table(read.csv(getFileName(2014), na=c("NULL", "PrivacySuppressed")))
```

```
subset(df, CITY == "Miami", c(INSTNM, ICLEVEL))
subset(df, CITY == "Miami", c(INSTNM, PREDDEG))
subset(df, CITY == "Miami" & INSTNM == "Florida Vocational Institute", c(INSTNM, CONTROL))
subset(df, CITY == "Miami" & INSTNM == "Florida International University", c(INSTNM, CONTROL))
subset(df, CITY == "Miami" & INSTNM == "Florida International University", c(INSTNM, SCH_DEG))
```

NOTE:

variable name	description	PageNo
UNK	revenues / expenses	8
CONTROL	0: public, 1: nonprofit, 2: profit	6
UNK	Retention Rates	14
UNK	Income	13
C200_4	20% Completioin 4-year	20
C200_l4	20% Completioin less-than-4-year	20
NPT4_PUB	Net price for public	11
NPT4_PRIV	Net price for private	11
NPT4i_*	Net price by quintile (i < 6)	11
ADM_RATE_ALL	Admission rate accross campuses	9
CCBASIC	Carnegie Classification	9

```
subset(df,TRUE, CIP01ASSOC)
subset(df,TRUE, DISTANCEONLY)
```

```
pcipIndex <- 62:99
fullModel <- lm(DEBT_MDN ~ ., data=subset(df, TRUE, c(1504,pcipIndex)))
stepped <- step(fullModel, direction = "backward")
```

```
summary(stepped)
```

```
fullPubModel <- lm(NPT4_PUB ~ ., data=subset(df, TRUE, c(which(names(df) == "NPT4_PUB"), pcipIndex)))
fullPrivModel <- lm(NPT4_PRIV ~ ., data=subset(df, TRUE, c(which(names(df) == "NPT4_PRIV"), pcipIndex))
pubStep <- step(fullPubModel, direction='backward')
privStep <- step(fullPrivModel, direction='backward')
```

```
summary(pubStep)
summary(privStep)
```

```
df$c200_total <- df$C200_4
df$c200_total[is.na(df$c200_total)] <- df$C200_L4[is.na(df$c200_total)]
df$c150_total <- df$C150_4
df$c150_total[is.na(df$c150_total)] <- df$C150_L4[is.na(df$c150_total)]
df$c100_total <- df$C100_4
df$c100_total[is.na(df$c100_total)] <- df$C100_L4[is.na(df$c100_total)]
```

The Carnegie rating and admission rates have pretty low R-Squared values in predicting completion rates.

```
tmp200Model <- lm(c200_total ~ CCBASIC + ADM_RATE, data=df)
tmp200Step <- step(tmp200Model, direction='backward')
summary(tmp200Step)
tmp150Model <- lm(c150_total ~ CCBASIC + ADM_RATE, data=df)
tmp150Step <- step(tmp150Model, direction='backward')
summary(tmp150Step)
tmp100Model <- lm(c100_total ~ CCBASIC + ADM_RATE, data=df)
tmp100Step <- step(tmp100Model, direction='backward')
summary(tmp100Step)
```

```
summary(pubStep)
summary(privStep)
```