QMSS Final Project

The Physician Compare website was created by the Centers for Medicare & Medicaid Services (CMS) in December 2010 as required by the Affordable Care Act (ACA) of 2010 to help patients assess and find doctors and hospitals. This dataset contains the information supplied to patients via that website, including patient satisfaction surveys and performance scores across over 100 metrics.

Looking at individual physician scores:

- MIPS
- Performance by measure category
- Organization MIPS

Possible project:
* Prediction of whether to see a physician based on threshold * Predict physician score based on all factors (predict MIPS?) + classification of whether you should go to a doctor based on MIPS score + train on several classification methods to analyze best approach for this data

Question/Problem: How can we better help patients assess and find doctors, where the scoring and rating come in a format not easily accessible or understandable by the average individual?

Approach/Methods: Supervised learning for binary classification utilizing the MIPS as a target with other physician scoring methods as predictors (which we know some of the metrics are direct factors of the individual MIPS scoring, such as the IA, ACI, and Quality category scorings). Potential methods outlined below, including generalized linear models and tree methods.

```
# Professional.Enrollment.ID is non-predictive, same with PAC_id
dr_scores <- read.csv("full_doctor_scoring.csv", sep = ",", na = c("NA", "N/A"))
```

Potential Methods for Binary Classification:
Using overall MIPS for individuals where MIPS>= 75, the positive payment adjustment threshold.

- could apply spline to other MIPS, ACI scorings since they're somewhat discrete in nature.
  - ACI >= 0 : clinician reported ACI category
  - ACI >= 50: clinician achieved base score for ACI
  - MIPS < 30: Negative Payment Adjustment
- Predictive MIPS >= 75, essentially.
- Methods to try:
  - glmnet for binary classification (elastic model/penalized logit)
  - glm logit model with polynomials?
  - tree model if we can make it work? (Single Tree, Random Forest, Boosting, Dbarts???)
  - PLSDA or LDA
  - nnet or MARS
- PCA to look at similar variables?

```
# remove majority of missingness by removing each task type -- though check how many are just preventat
no_tasks <- dr_scores[c(1:23, 57)]

# add a "UKN" for unknown category
no_tasks$Credential <- factor(no_tasks$Credential, levels = c(levels(dr_scores$Credential), "UKN"))
```

```r
no_tasks$Credential[is.na(no_tasks$Credential)] <- "UKN"

# -1 for not reporting individual ACI Score (NA value)
no_tasks$ACI_Score_Dr[is.na(no_tasks$ACI_Score_Dr)] <- -1

head(no_tasks)
```

```
##       PAC_id Professional.Enrollment.ID last_name first_name Gender
## 1 4385734086            I20071219000090    KANTER      DAVID      F
## 2 7214189315            I20121219000307  GOTESMAN  ALEXANDER      M
## 3 6507956612            I20071213000113   HARTMAN    MATTHEW      M
## 4 8325134752            I20110720000797 NASAJPOUR    HOSSEIN      M
## 5 8325216575            I20180518001359  MCCOPPIN      HOLLY      F
## 6 9133275274            I20090915000481      JOHN  JAYASHREE      F
##   Credential                                  Medical.school.name
## 1        UKN                                                OTHER
## 2        UKN                                                OTHER
## 3        UKN         UNIVERSITY OF PITTSBURGH SCHOOL OF MEDICINE
## 4        UKN           UNIVERSITY OF FLORIDA COLLEGE OF MEDICINE
## 5        UKN UNIVERSITY OF MISSOURI, COLUMBIA SCHOOL OF MEDICINE
## 6        UKN                                                OTHER
##   Graduation.year                    Primary.specialty Num.Secondaries
## 1            1998 PHYSICAL MEDICINE AND REHABILITATION               0
## 2            1999                              UROLOGY               0
## 3            2002                 DIAGNOSTIC RADIOLOGY               0
## 4            2003   PLASTIC AND RECONSTRUCTIVE SURGERY               0
## 5            2006                           DERMATOLOGY               0
## 6            1984                            PSYCHIATRY               0
##   Group.Practice.PAC.ID State
## 1            8123911500    NY
## 2            9032293709    NJ
## 3            8426364738    PA
## 4            4385876655    MS
## 5            1456329655    CO
## 6            8729381033    GA
##                            Organization.legal.name Score_Source_Org
## 1 PHYSICAL MEDICINE and REHAB MEDICAL SERVICE GROUP            group
## 2        LAKEWOOD UROLOGY LIMITED LIABILITY COMPANY            group
## 3                       ALLEGHENY CLINIC RADIOLOGY            group
## 4                       SOUTH CENTRAL CLINICS, INC            group
## 5                    LAKE LOVELAND DERMATOLOGY PC             group
## 6                ST FRANCIS PHYSICIAN PRACTICES LLC            group
##   Quality_Score_Org ACI_Score_Org IA_Score_Org MIPS_Score_Org
## 1               100           100           40         100.00
## 2                 0             0            0           0.00
## 3               100             0           40         100.00
## 4                 0             0           20           7.50
## 5                 0             0            0           0.00
## 6                 0             0           10           3.75
##   Score_Source_Dr Quality_Score_Dr ACI_Score_Dr IA_Score_Dr MIPS_Score_Dr
## 1           group            100.0          100          40         100.0
## 2      individual             67.0            0           0          40.2
## 3           group            100.0           -1          40         100.0
```

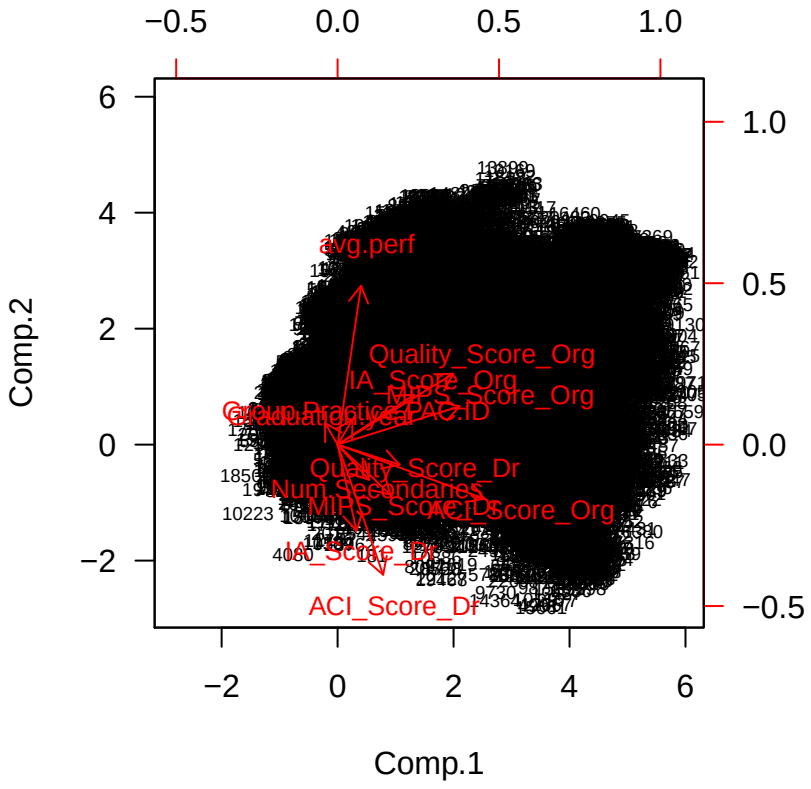
```
## 4      individual             69.2          100           40           81.5
## 5      individual             72.6          100           40           83.6
## 6      individual             22.9           -1            0           19.5
##   avg.perf
## 1     74.0
## 2     87.6
## 3    100.0
## 4     47.0
## 5     54.4
## 6     99.0
```

PCA to Explore Correlation of Variables

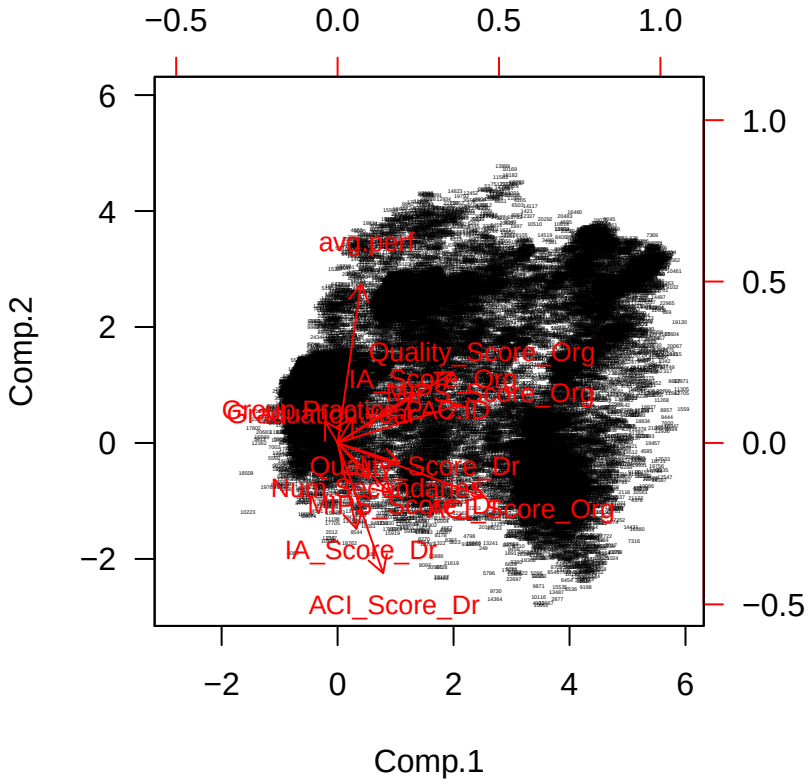
An initial PCA analysis to look understand how rows cluster based on the column variables. This is without considering the binary classifying aspect of MIPS >= 75„ and is just to understand some of the structure of the scores in the data.

```
# still minor missingness in Graduation Year and Quality Score
numerics <- no_tasks[, c(8, 10:11, 15:18, 20:24)] %>% na.omit(.)

pr_out <- PCAproj(numerics, scale = sd)
par(mar = c(5, 4, 3, 3) + 0.1, las = 1)
biplot(pr_out, scale = 0, cex = c(0.6, 0.8)) # visualization of any immediate outliers and the variable
```



```r
biplot(pr_out, scale = 0, cex = c(0.13, 0.8)) # clusters within the observations visible
```



As mostly expected, the MIPS scores for the hospital organizations that each doctor works at are more correlated to each other than they are to the MIPS scores for each individual doctor. While the number of secondary specialties each doctor has is more correlated to the individual doctor MIPS scores, the group practice ID is more correlated to the organization scores. This is also fairly in line with our expectations that measures for the practice organization would cluster separately from the measures for the individual doctors. Interestingly, the average performance for an individual doctor across task categories seems to be more correlated to the organization scores though. Since performance is measured though individual patient reporting, their experience with the organization itself may be taken into consideration and bias their score, even if the doctor-patient interaction itself was positive.

In the second PCA plot, with the observation labels less cluttered from size, we see that there seem to be distinct clusters of observations. This would indicate groupings within the observations with distinguishing measurement characteristics. Many points do deviate from the groupings themselves. Still, overall, it doesn't appear that any observations seems like a major outlier, as seen in the first plot.