Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

Eric Yang, Samir Datta, Carlos Castro

oCTOBER 22, 2017

Introduction

Here we present the results of our analysis on contributions data for the university foundation, where our goal is to utilize the data available to predict who are likely to donate in the future and an idea of the magnitude of such donation.

The dataset includes, for each record, information such as:

- Donation amounts for the last 4 years
- Whether they attended the contribution events between 2012 and 2015
- Year of graduation
- Marital status, Gender
- Major of studies
- Year of graduation

Given that the goal of this study is to be able to predict who are likely to donate and the magnitude of that donation, we build a model focusing on its predictive power rather than its explanatory power. To build our model, in the following sections we thoroughly analyze the data, afterwards consider both multinomial and ordinal models and conduct statistical analysis on them to choose the one most fit for the required predictions.

<TODO: Concise summary of results here!>

Section 2

```
#loading packages and data
library(ggplot2)
library(ordinal)
```

```
## Warning: package 'ordinal' was built under R version 3.4.2
```

```
library(nnet)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.2
```

```
## Loading required package: lattice
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.4.2
```

```
labdata <- read.csv('C:/Users/Samir/Documents/MIDS/StatsF17/lab 2/lab2data.csv')
```

Feature enginerring

```
#split major into STEM and non-STEM
labdata$MajorType <- ifelse(labdata$Major=='Biology'|
                              labdata$Major=='Economics'|
                              labdata$Major=='Psychology'|
                              labdata$Major=='Physics'|
                              labdata$Major=='Chemistry'|
                              labdata$Major=='Mathematics'|
                              labdata$Major=='General Science-Chemistry'|
                              labdata$Major=='Economics-Business'|
                              labdata$Major=='General Science-Chemistry'|
                              labdata$Major=='Sociology-Anthropology'|
                              labdata$Major=='General Science-Psycho'|
                              labdata$Major=='General Science-Math'|
                              labdata$Major=='General Science-Biology'|
                              labdata$Major=='Computer Science'|
                              labdata$Major=='General Science'|
                              labdata$Major=='Mathematics-Physics'|
                              labdata$Major=='Economics-Regional Stds.'|
                              labdata$Major=='Zoology'|
                              labdata$Major=='Engineering'|
                              labdata$Major=='Sociology'|
                              labdata$Major=='Anthropology'|
                              labdata$Major=='General Science-Physics',
                            "STEM", "Non-STEM")

#create variable nextDegreeType to categorize the most common next degrees
labdata$NextDegreeType <- ifelse(labdata$Next.Degree=='JD', 'JD',
                          ifelse(labdata$Next.Degree=='MA', 'MA',
                          ifelse(labdata$Next.Degree=='PHD', 'PHD',
                         ifelse(labdata$Next.Degree=='NDA', 'NDA',
                                ifelse(labdata$Next.Degree=='MS', 'MS',
                                       ifelse(labdata$Next.Degree=='MD', 'MD',
                                              ifelse(labdata$Next.Degree=='MBA', 'MBA',
        ifelse(labdata$Next.Degree=='NONE', 'NONE', 'Other'))))))))

#create simpler variable to represent if someone has an advanced degree or not
labdata$NextDegreeBinary <- ifelse(labdata$Next.Degree=='NONE', 0, 1)


#create buckets for all years
labdata$FY16cat <- cut(labdata$FY16Giving, c(0,1,100,250,500,200000), right=F)
labdata$FY15cat <- cut(labdata$FY15Giving, c(0,1,100,250,500,200000), right=F)
labdata$FY14cat <- cut(labdata$FY14Giving, c(0,1,100,250,500,200000), right=F)
labdata$FY13cat <- cut(labdata$FY13Giving, c(0,1,100,250,500,200000), right=F)
labdata$FY12cat <- cut(labdata$FY12Giving, c(0,1,100,250,500,200000), right=F)

#turn class year into years since grad to make interpretaion easier
labdata$YearsSinceGrad <- 2017 - labdata$Class.Year


#loop through data to get each person's mean donation over the past 4 years
```

```
#and how many of the past years they've donated
labdata$meandonation <- NA
labdata$nPastYears <- NA
for (i in c(1:1000)){
  labdata[i,]$meandonation<-mean(c(labdata[i,]$FY12Giving,
                                   labdata[i,]$FY13Giving,
                                   labdata[i,]$FY14Giving,
                                   labdata[i,]$FY15Giving))

  labdata[i,]$nPastYears <- sum(c(labdata[i,]$FY12Giving>0,
                                  labdata[i,]$FY13Giving>0,
                                  labdata[i,]$FY14Giving>0,
                                  labdata[i,]$FY15Giving>0))
}

#binary variable - have they donated before or not?
labdata$past_binary <- ifelse(labdata$meandonation == 0,0,1)

#did they donate last year?
labdata$donatedLastYear <- ifelse(labdata$FY15Giving==0, 0, 1)

#how many of the past 4 years, consecutively, did they donate?
#note that this will give a 0 for those that donated 2012-2014 but NOT 2015
#since we're asking for consecutive years
labdata$nPastConsecutiveYears <- ifelse(labdata$FY15Giving==0, 0,
                                        ifelse(labdata$FY14Giving==0,1,
                                               ifelse(labdata$FY13Giving==0,2,
                                                      ifelse(labdata$FY12Giving==0,3,4))))

#did they give in 2015 or not?
labdata$gaveLastYear <- ifelse(labdata$FY15Giving==0,0,1)
```

```
labdata$donatedLastYear <- ifelse(labdata$FY15Giving==0, 0, 1)

labdata$nPastConsecutiveYears <- ifelse(labdata$FY15Giving==0, 0,
                                        ifelse(labdata$FY14Giving==0,1,
                                               ifelse(labdata$FY13Giving==0,2,
                                                      ifelse(labdata$FY12Giving==0,3,4))))
labdata$gaveLastYear <- ifelse(labdata$FY15Giving==0,0,1)
```

EDA

Univariate analysis

Gender

```
table(labdata$Gender)
```

```
##
##   F   M
## 505 495
```

Both male and female alumni are approximately evenly represented.

Years since graduation

```
table(labdata$YearsSinceGrad)
```

```
##
##   5  15  25  35  45
## 293 223 203 176 105
```

For the sake of an easier interpretation we transformed the varaible "Class.Year" into years since graduation by subtracting it from 2017. There are 5 unique values which reveals that this dataset polled alumni from classes 10 years apart. Younger alumni are more represented - alumni who graduated 5 years ago are represented almost 3 times as much as those who graduated 45 years ago. Already this appears to be an important variable to control for so that our model is not biased towards younger graduates. Furthermoer, due to the groups of the variable, we will want to treat this as an ordinal - instead of a continuous - variable.

Marital.Status

```
table(labdata$Marital.Status)
```

```
##
##   D   M   S   W
##  61 584 344  11
```

Most alumni are married, a good portion are single as well. The divorced and widowed group are sparsely represented.

Major

```
length(unique(labdata$Major))
```

```
## [1] 45
```

```
table(labdata$MajorType)
```

```
##
## Non-STEM     STEM
##      522      478
```

There are 45 different majors with varying levels of representation, including many with only one alumnus (for the sake of saving space we have decided to not show the full list). Because of that, we condensed this variable into STEM vs. Non-STEM, both of which appear to be approximately equally represented.

Next Degree

```
table(labdata$Next.Degree)
```

```
##
##   AA   BA  BAE   BD  BFA   BN   BS  BSN   DC  DDS  DMD   DO  DO2   DP   JD
##    1    4    1    1    1    2    2    3    1    1    1    2    1    1   90
##  LLB  LLD   MA  MA2  MAE MALS  MAT  MBA  MCP   MD  MD2   ME  MFA  MHA   ML
##    1    1  108    1    1    1   10   34    1   42    9   17   14    1    1
```

```
##  MLS   MM  MPA  MPH   MS  MSM  MSW  NDA NONE  PHD  STM   TC UBDS UDDS  UMD
##    9    1    6    4   53    1   11   58  378   78    1   22    6    4    6
## UMDS UNKD
##    2    6
```

```
table(labdata$NextDegreeBinary)
```

```
##
##   0   1
## 378 622
```

Like the Major variable, there is a variety of sparsely represented advanced degrees, so we chose to condense it into a binary variable - "None" vs. the rest. Interestingly, a considerable majority of alumni in this sample have an advanced degree, which could point to a sampling bias.

Attendance Event

```
table(labdata$AttendenceEvent)
```

```
##
##   0   1
## 395 605
```

A majority of alumni have attended alumni events between 2012 and 2015. This could also point to sampling bias - the dataset may come from alumni who were already more likely to donate than not.

Previous donations

```
table(labdata$FY12cat)
```

```
##
##       [0,1)     [1,100)   [100,250)   [250,500) [500,2e+05)
##         558         213         149          37          43
```

```
table(labdata$FY13cat)
```

```
##
##       [0,1)     [1,100)   [100,250)   [250,500) [500,2e+05)
##         513         247         143          54          43
```

```
table(labdata$FY14cat)
```

```
##
##       [0,1)     [1,100)   [100,250)   [250,500) [500,2e+05)
##         553         226         136          36          49
```

```
table(labdata$FY15cat)
```

```
##
##       [0,1)     [1,100)   [100,250)   [250,500) [500,2e+05)
##         567         199         138          36          60
```

From 2012 to 2015 the number of people in each donation category appears relativly stable. Higher donation categories have less alumni, with the exception of the highest category [500,2e+05) which has more than the next highest one in 3/4 years.

FY16 category

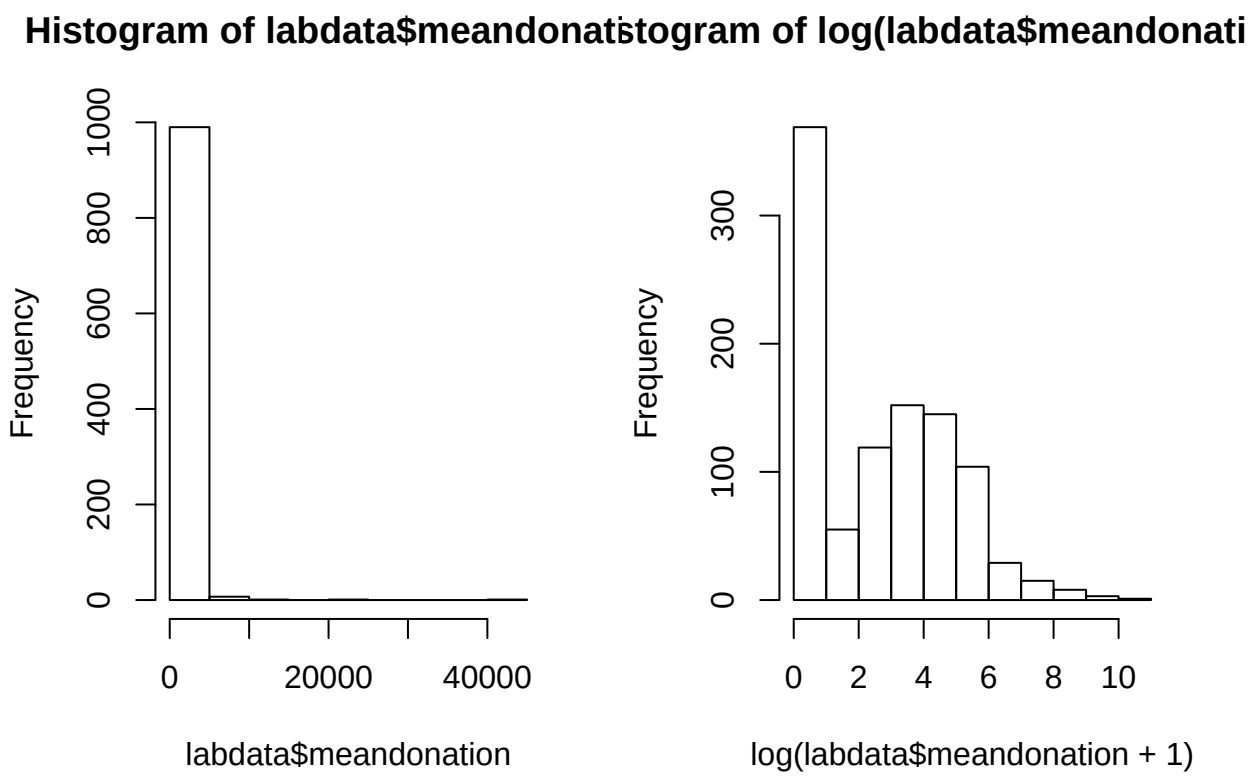
```
table(labdata$FY16cat)
```

```
##
##       [0,1)     [1,100)   [100,250)   [250,500) [500,2e+05)
##         586         173         143          39          59
```

The numbers for 2016 also look very similar to the previous years. This suggests that a large number of alumni stay in the same donation category from year to year, and that implementing information about previous years' donations will be crucial for our model's predictive ability. As we noticed before, the [250,500) category is very sparsely represented, which may make it hard to predict accurately.

mean donation in the past

```
par(mfrow=c(1,2))
hist(labdata$meandonation)
hist(log(labdata$meandonation+1))
```

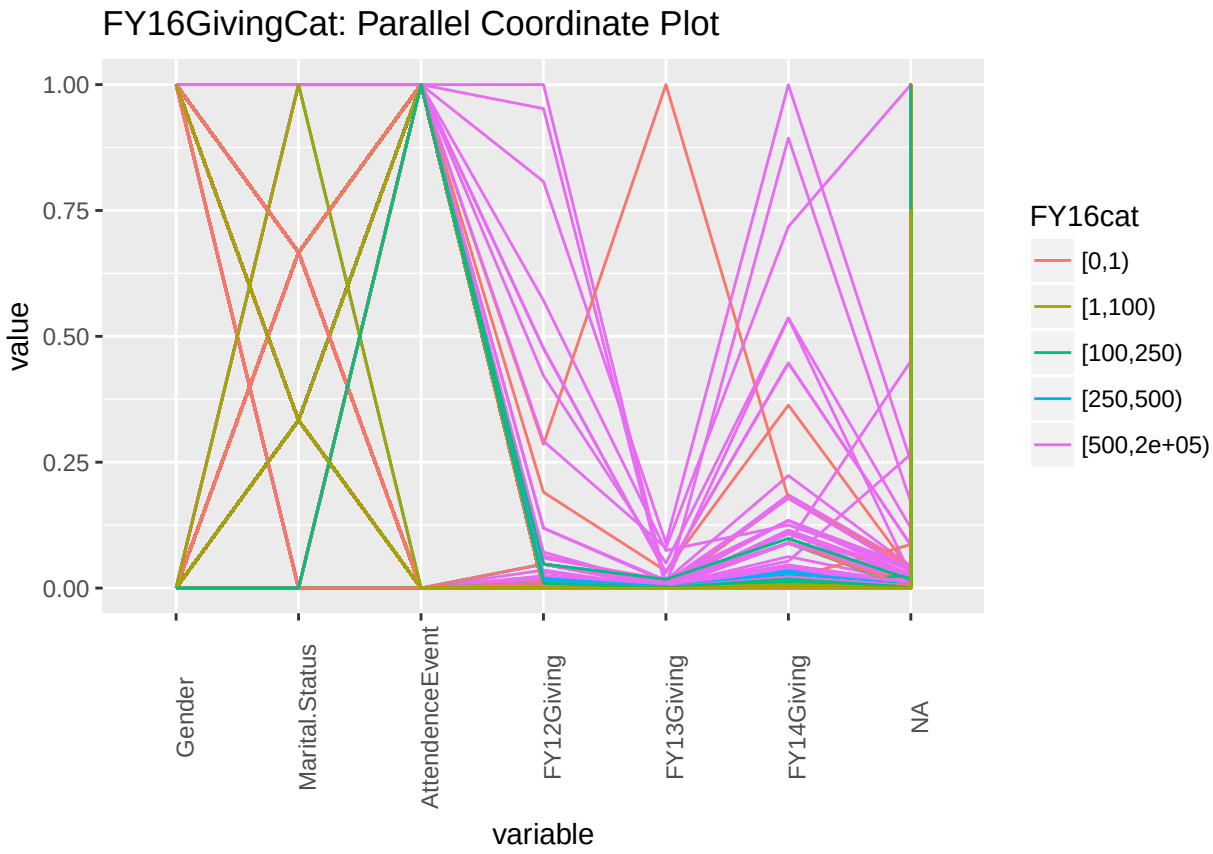


The variable mean donation, which represents each alumnus' mean donation from 2012-2015, has a large positive skew. Applying a log transformation (after adding 1, since the value 0 can't be log transformed) solves this to some extent, although a slight positive skew is still evident. A disproportionate number of alumni have a mean donation value of 0.

Relationship between FY16cat and other variables

Parallel coordinate plot

```
parallel.columns <- c("Gender", "Marital.Status", "AttendenceEvent","FY12Giving", "FY13Giving", "FY14Gi
ggparcoord(labdata, columns = parallel.columns, order = 1:length(parallel.columns), groupColumn = "FY16
```

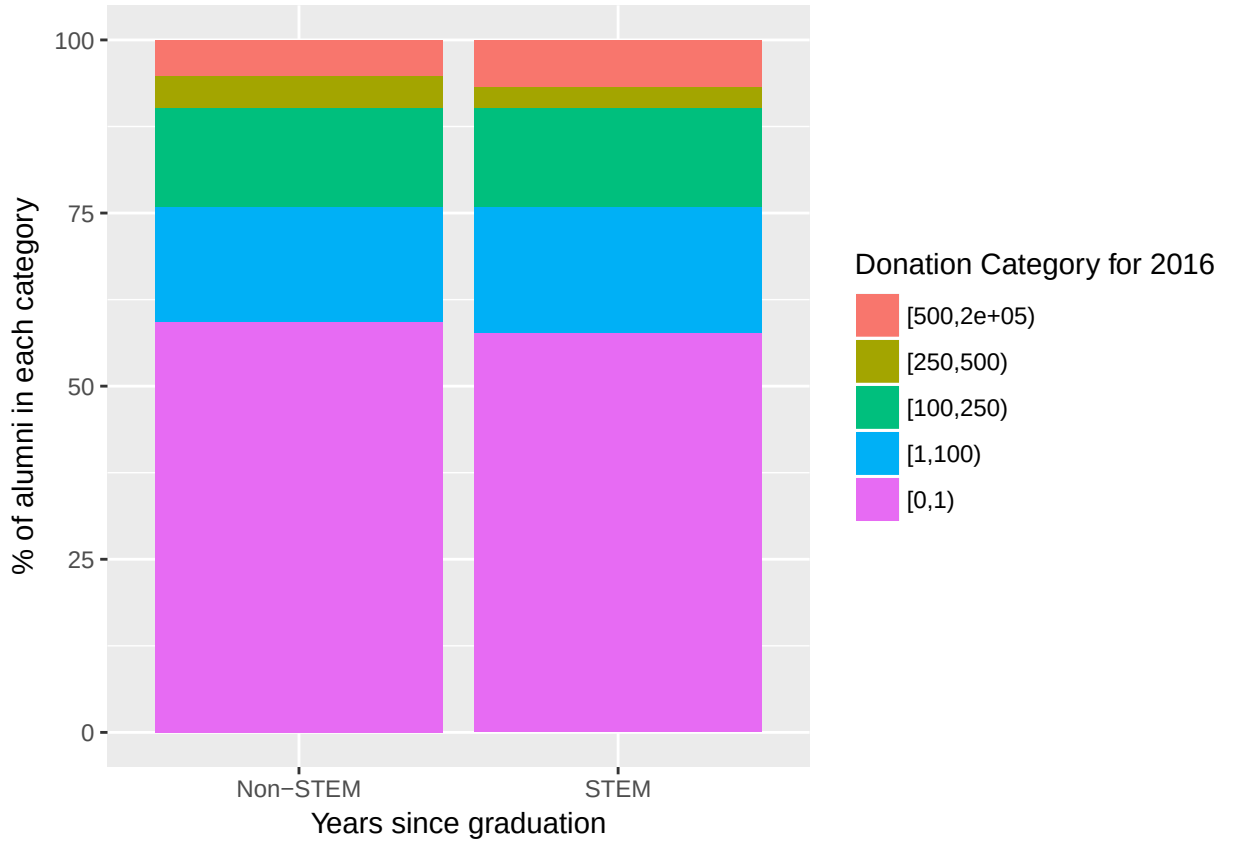


Major type - STEM vs. Non-STEM

```
labdata_counts <- with(labdata,
                       aggregate(MajorType,
                                 list(MajorType=MajorType),
                                 length))
labdata_agg <- with(labdata,
                    aggregate(MajorType, list(MajorType=MajorType,
                                              FY16cat=FY16cat),
                    length))
labdata_agg <- merge(labdata_agg, labdata_counts, by="MajorType")
labdata_agg <- setNames(labdata_agg, c("MajorType", "FY16cat", "Count", "TotalCount"))
labdata_agg$percent <- 100*labdata_agg$Count/labdata_agg$TotalCount
labdata_agg$FY16cat <- factor(labdata_agg$FY16cat,
      levels=c("[500,2e+05)","[250,500)","[100,250)","[1,100)","[0,1)"))

ggp <- ggplot(labdata_agg, aes(x=MajorType,y=percent))
```

```
ggp + geom_bar(stat="identity", aes(fill=FY16cat))+
  ylab("% of alumni in each category")+xlab("Years since graduation")+
  scale_fill_discrete(name="Donation Category for 2016")
```



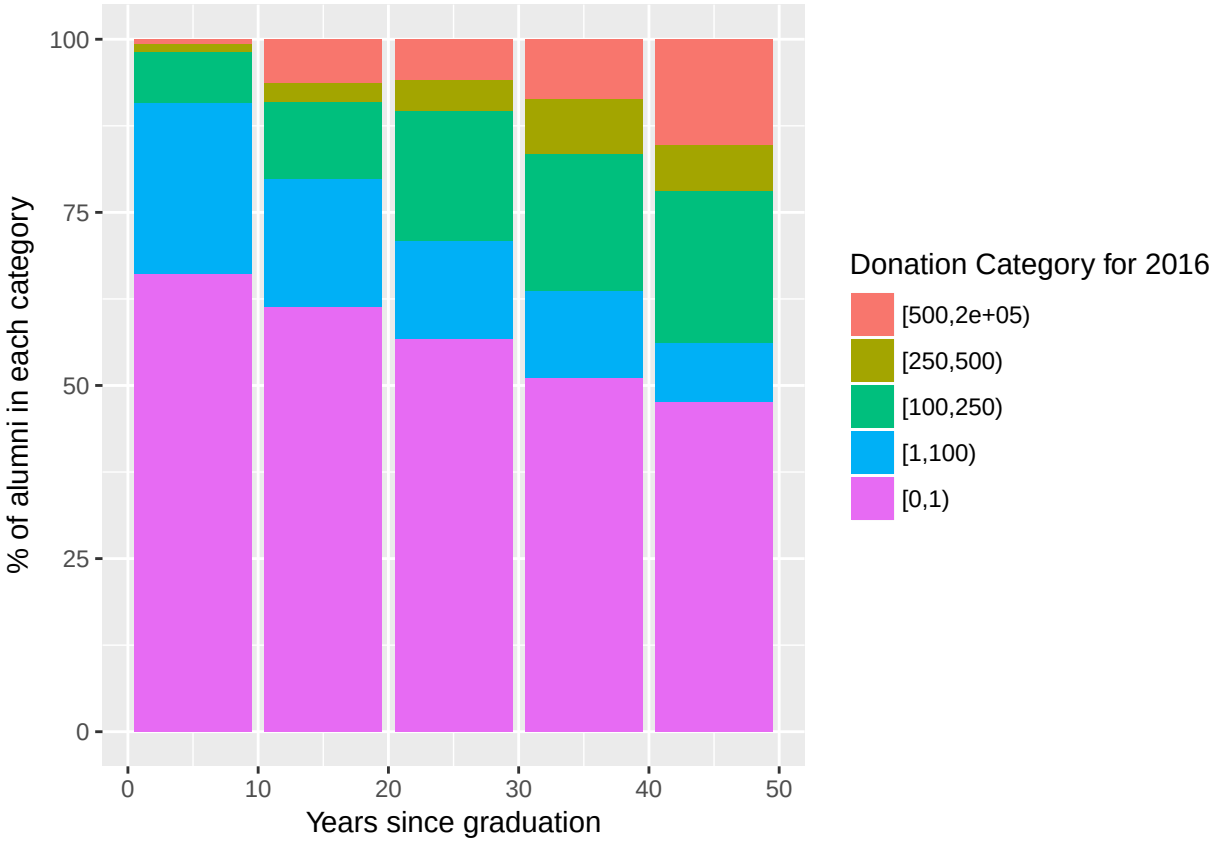
Major type does not seem to have an effect on the donation amount of an alumnnus, as the distribution of donation categories appears virtually identical regardless of whether they graduated with a STEM or non-STEM degree.

Next Degree (binary)

```
labdata_counts <- with(labdata,
                       aggregate(YearsSinceGrad,
                                 list(YearsSinceGrad=YearsSinceGrad),
                                 length))
labdata_agg <- with(labdata,
                    aggregate(YearsSinceGrad, list(YearsSinceGrad=YearsSinceGrad,
                                                   FY16cat=FY16cat),
                    length))
labdata_agg <- merge(labdata_agg, labdata_counts, by="YearsSinceGrad")
labdata_agg <- setNames(labdata_agg, c("YearsSinceGrad", "FY16cat", "Count", "TotalCount"))
labdata_agg$percent <- 100*labdata_agg$Count/labdata_agg$TotalCount
labdata_agg$FY16cat <- factor(labdata_agg$FY16cat,
      levels=c("[500,2e+05)","[250,500)","[100,250)","[1,100)","[0,1)"))
```

```
ggp <- ggplot(labdata_agg, aes(x=YearsSinceGrad,y=percent))

ggp + geom_bar(stat="identity", aes(fill=FY16cat))+
  ylab("% of alumni in each category")+xlab("Years since graduation")+
  scale_fill_discrete(name="Donation Category for 2016")
```



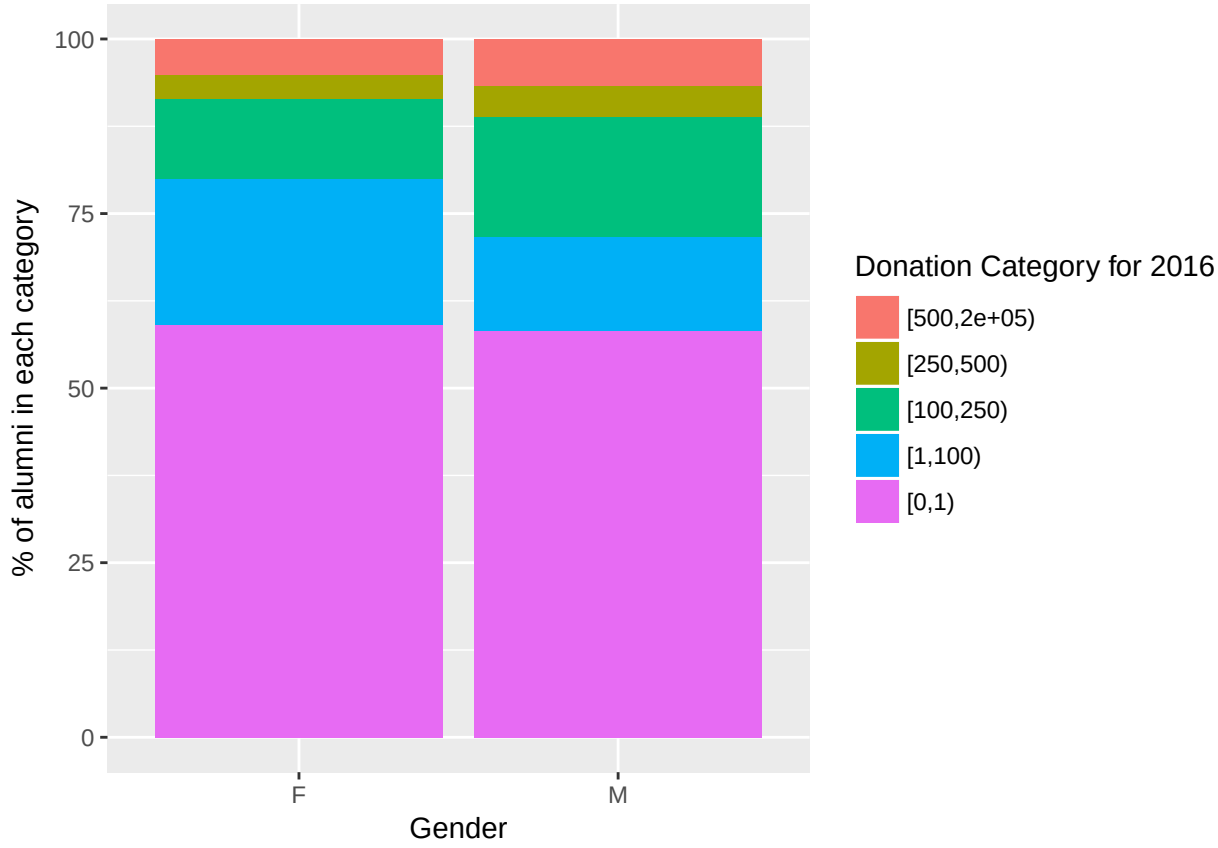
A clear ordinal relationship between years since grad and the amount donated in 2016 is shown in the violin plot, where those that graduated longer ago are more likely to not be in the [0,1] category and more likely to be in higher donation categories as well.

Gender

```
labdata_counts <- with(labdata,
                       aggregate(Gender,
                                 list(Gender=Gender),
                                 length))
labdata_agg <- with(labdata,
                    aggregate(Gender, list(Gender=Gender,
                                           FY16cat=FY16cat),
                    length))
labdata_agg <- merge(labdata_agg, labdata_counts, by="Gender")
labdata_agg <- setNames(labdata_agg, c("Gender", "FY16cat", "Count", "TotalCount"))
labdata_agg$percent <- 100*labdata_agg$Count/labdata_agg$TotalCount
labdata_agg$FY16cat <- factor(labdata_agg$FY16cat,
      levels=c("[500,2e+05)","[250,500)","[100,250)","[1,100)","[0,1)"))
```

```
ggp <- ggplot(labdata_agg, aes(x=Gender,y=percent))

ggp + geom_bar(stat="identity", aes(fill=FY16cat))+
  ylab("% of alumni in each category")+xlab("Gender")+
  scale_fill_discrete(name="Donation Category for 2016")
```



Men appear to be more likely to donate in the top 3 categories, while women appear to be more likely to donate in the [1,100) category. Interestingly, both men and women appear to be just as likely to donate nothing. This may suggest that gender would be more useful for a multinomial model instead of an ordinal model.

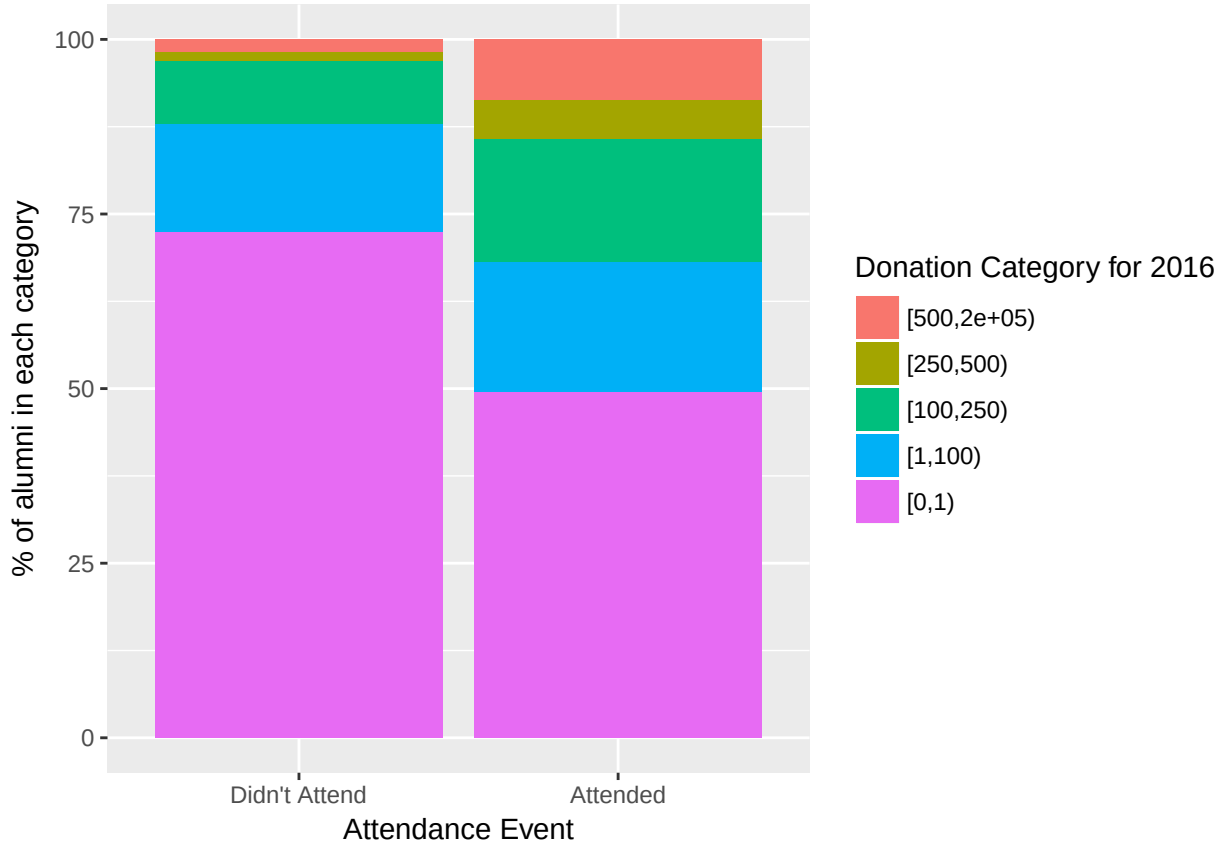
Attendance event

```
labdata_counts <- with(labdata,
                       aggregate(AttendenceEvent,
                                 list(AttendenceEvent=AttendenceEvent),
                                 length))
labdata_agg <- with(labdata,
                    aggregate(AttendenceEvent, list(AttendenceEvent=AttendenceEvent,
                                                    FY16cat=FY16cat),
                    length))
labdata_agg <- merge(labdata_agg, labdata_counts, by="AttendenceEvent")
labdata_agg <- setNames(labdata_agg, c("AttendenceEvent", "FY16cat", "Count", "TotalCount"))
labdata_agg$percent <- 100*labdata_agg$Count/labdata_agg$TotalCount
```

```
labdata_agg$FY16cat <- factor(labdata_agg$FY16cat,
      levels=c("[500,2e+05)","[250,500)","[100,250)","[1,100)","[0,1)"))

ggp <- ggplot(labdata_agg, aes(x=as.factor(AttendenceEvent),y=percent))

ggp + geom_bar(stat="identity", aes(fill=FY16cat))+
  ylab("% of alumni in each category")+xlab("Donation category for 2015")+
  scale_fill_discrete(name="Donation Category for 2016")+
  xlab("Attendance Event")+scale_x_discrete(labels=c("Didn't Attend", "Attended"))
```



Those who went to alumni events were much more likely to donate and especially more likely to donate in the higher categories.

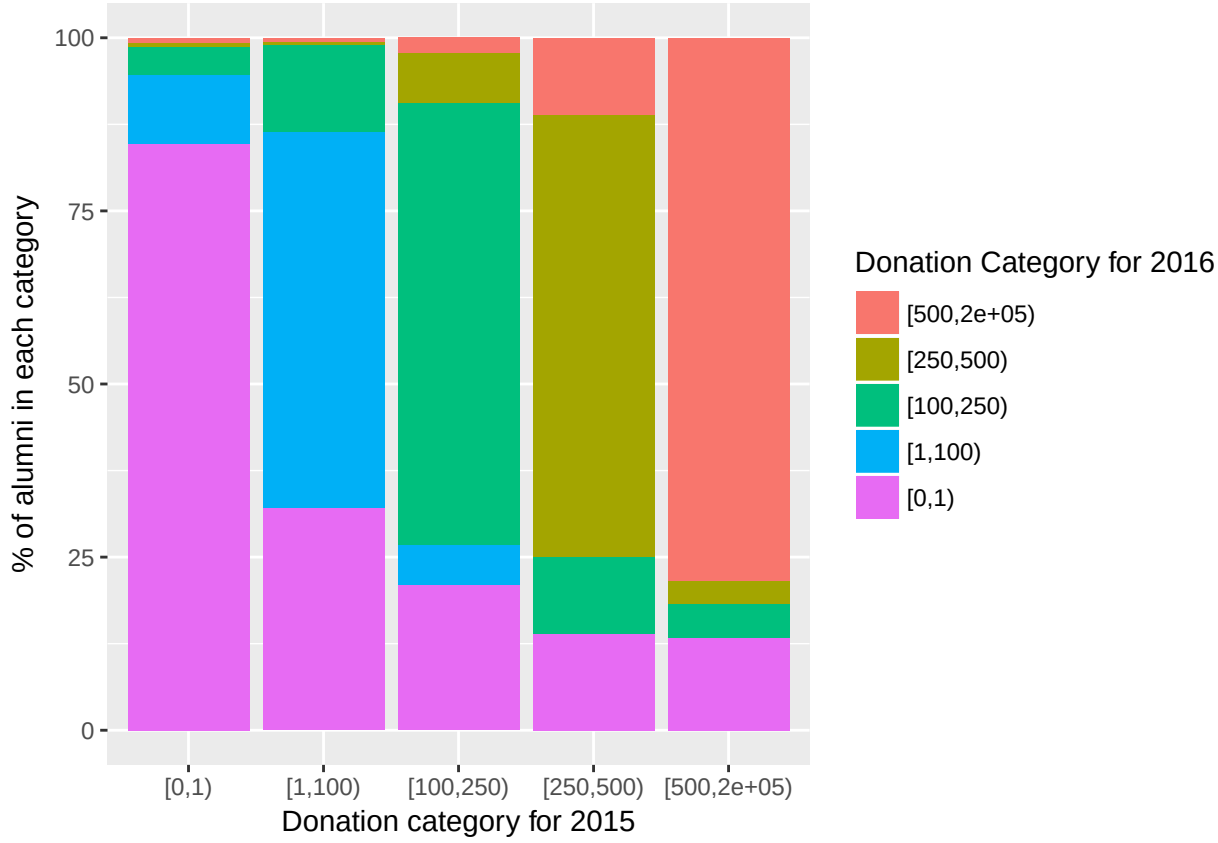
Donation category for the previous year (2015)

```
labdata_counts <- with(labdata,
                       aggregate(FY15cat,
                                 list(FY15cat=FY15cat),
                                 length))
labdata_agg <- with(labdata,
                    aggregate(FY15cat, list(FY15cat=FY15cat,
                                            FY16cat=FY16cat),
                    length))
labdata_agg <- merge(labdata_agg, labdata_counts, by="FY15cat")
labdata_agg <- setNames(labdata_agg, c("FY15cat", "FY16cat", "Count", "TotalCount"))
```

```
labdata_agg$percent <- 100*labdata_agg$Count/labdata_agg$TotalCount
labdata_agg$FY16cat <- factor(labdata_agg$FY16cat,
      levels=c("[500,2e+05)","[250,500)","[100,250)","[1,100)","[0,1)"))

ggp <- ggplot(labdata_agg, aes(x=FY15cat,y=percent))

ggp + geom_bar(stat="identity", aes(fill=FY16cat))+
  ylab("% of alumni in each category")+xlab("Donation category for 2015")+
  scale_fill_discrete(name="Donation Category for 2016")
```



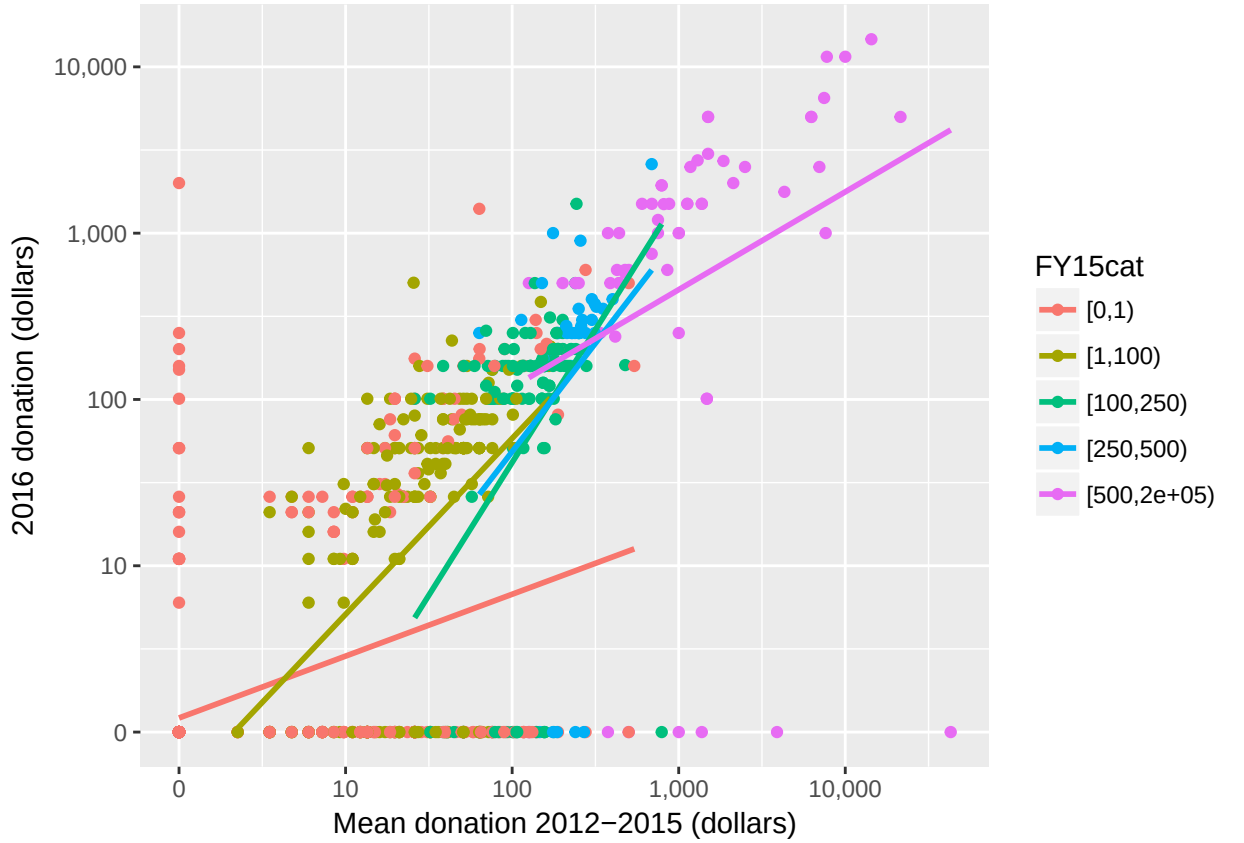
This stacked bar graph shows what proportion of alumni that fit into donation category X went into the same - or different - category in 2016. As expected, the largest bar in each group represents the same category. That is, the majority of those who donated $0 in 2015 also donated $0 in 2016, the majority of those who donated between $1-$100 in 2015 stayed in that category the next year, etc.

A key takeaway from this visualization is the relative instability of the [1,100) class - compared to other classes, this group was the least likely to donate in the same category. Still, past donations seem to be important in predicting future donations.

Interaction between last year's donations and overall donations in predicting 2016 donations

```
ggp <- ggplot(labdata, aes(x=log(meandonation+1)/log(10), y=log(FY16Giving+1)/log(10),
                           group=FY15cat, color=FY15cat))
```

```
ggp + geom_point() + geom_smooth(method="lm", se=F)+
  xlab("Mean donation 2012-2015 (dollars)")+ylab("2016 donation (dollars)")+
  scale_x_continuous(breaks=c(0,1,2,3,4), labels=c("0", "10", "100", "1,000", "10,000"))+
    scale_y_continuous(breaks=c(0,1,2,3,4), labels=c("0", "10", "100", "1,000", "10,000"))
```



Above is a scatterplot with the mean donation from 2012-2015 on the x-axis and the amount donated in 2016 on the y-axis. (While we are analyzing 2016 donations in categories, we thought this visualization was best done with the dollar amount). Overall, there is a clear relationship between mean donation and amount donated in 2016 - alumni typically didn't donate a drastically different amount in 2016 compared to how they've donated in years past. Of course, the exceptions are the many alumni who didn't donate in 2016 despite donating in years past (the dots on the horizontal x=0 line). There are a lot less alumni who donated in 2016 for the first time (the dots on the vertical y=0 line)

The purpose of the different colors/trend lines is to examine an interaction effect we found interesting and potentially useful for our model. The lines seem to generally be parallel except for the one representing the [0,1) category. The implication of this is that for alumni who donated in 2015, it is easier to predict their 2016 donations from their previous years, but for alumni who did not donate in 2015 the relationship is less clear. Rather than modeling an interaction term for each category for 2015, which could get too complex, it seems the interaction comes from whether they donated at all last year or not, which is why we will model their 2015 donations as a binary variable.

Section 3

Section 4