W271 Section 3 Lab 2

Kiersten Henderson, Jill Zhang, Hoang Phan, Daghan Altas

10/8/2017

```
library(knitr)
library(vcd)
opts_chunk$set(tidy.opts=list(width.cutoff=75),tidy=TRUE)
library(Hmisc)
library(ggplot2)
library(dplyr)
library(GGally)
library(data.table)
library(stargazer)
library(tidyverse)
library(forcats)
library(scales)
library(gridExtra)
```

1 Introduction

Our team of data scientists is working with a University foundation to identify alumni who are likely to donate in the future. We were able to model not just whether alumni with particular characteristics would donate, but rather what level of contribution particular alumni were most likely to make.

In order to evaluate what variables to include when we estimated our proportional odds model, we considered fundraising domain knowledge and the results of our explanatory data analysis (EDA). Fundraising professionals typically consider an individual's wealth (ability to give) and enthusiasm for giving when they consider who might be the most likely, loyal, and generous donors to a particular cause. For this reason and due to our EDA, we included Class.Year in the model we estimated because it is a proxy for wealth (in our dataset, the alumni who have graduated the longest ago likely have the highest earning potential). Similarly, we included Next.Degree as another potential indicator of wealth (those individuals with graduate degrees frequently have higher incomes). Enthusiasm for and support of the university is evident if an alumnus has attended a university event, and also if they have previously donated to their alma matter. For these reasons and due to our EDA, we included Event.Attendence and all previous years donations to the university in our model (FY12Giving-FY16Giving). In addition, we found in our EDA that Gender played a role in the philanthropic behavior towards the University and we thus included Gender when we estimated our model.

describe key results of model

2 Exploratory Data Analysis

```
givings = read.csv("./lab2data.csv")
str(givings)
```

```
## 'data.frame':    1000 obs. of  12 variables:
##  $ X              : int  761 620 214 373 748 1080 1155 1069 1161 457 ...
##  $ Gender         : Factor w/ 2 levels "F","M": 1 2 1 1 2 1 1 1 1 1 ...
##  $ Class.Year     : int  2002 2002 1982 1992 2002 2012 2012 2012 2012 1992 ...
```

```
##  $ Marital.Status : Factor w/ 4 levels "D","M","S","W": 2 3 2 2 3 3 3 3 3 2 ...
##  $ Major          : Factor w/ 45 levels "American Studies",..: 39 25 25 2 30 2 3 26 39 15 ...
##  $ Next.Degree    : Factor w/ 47 levels "AA","BA","BAE",..: 37 39 39 35 39 15 39 35 39 18 ...
##  $ AttendenceEvent: int  1 0 1 1 0 1 0 1 0 0 ...
##  $ FY12Giving     : num  50 0 100 0 0 0 0 5 0 0 ...
##  $ FY13Giving     : num  51 0 0 0 0 0 0 10 0 75 ...
##  $ FY14Giving     : num  51 0 100 0 0 0 0 25 0 0 ...
##  $ FY15Giving     : num  0 0 100 0 0 0 0 25 0 0 ...
##  $ FY16Giving     : num  0 0 100 0 0 0 0 50 0 60 ...
```

```
sum(is.na(givings))
```

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

2.1 Observations

- There are no missing variables, which simplifies the data clean-up task.
- There are 1000 observations and twelve variables (five of them are donations in different years).
- FY2016 is the dependent variable that we'd like to predict. However, we are given FYGiving for years 2012 through 2016 as amount in dollars (a c4ontinuous variable).
- Maximum donation is $161500 (in 2013)
- Gender is a binary variable.
- Marital status has four categories (D, M, S, W), which we interpret as divorced, married, single, windowed.
- Graduating class is strangely in five categories each ten years apart (1972, 1982, 1992, 2002, 2012).
- Donor's major is a categorical variable with 45 categories.
- Attendance of events is a binary category variable (0 for no, 1 for yes).
- Next degree is a categorical variable with 47 categories.

After i cleaned up the variables, i need to go back and describe each one (univariate analysis).

2.2 Data clean-up

First, we are going to clean-up the factor variables by providing explicit values for each level.

```
levels(givings$Gender) = c("Female", "Male")
givings$AttendenceEvent = factor(givings$AttendenceEvent, levels = c(0, 1),
    labels = c("Didn't Attend", "Attended"))
levels(givings$Marital.Status) = c("Divorced", "Married", "Single", "Widowed")
givings$Class.Year = factor(givings$Class.Year)
givings$FY12Giving = as.numeric(givings$FY12Giving)
givings$FY13Giving = as.numeric(givings$FY13Giving)
givings$FY14Giving = as.numeric(givings$FY14Giving)
givings$FY15Giving = as.numeric(givings$FY15Giving)
givings$FY16Giving = as.numeric(givings$FY16Giving)
```

We are going to create factor variables out of donations, since we are asked to group FY2016 donations to 5 buckets, we have decided to apply that same logic to all other years.

```
givings$FY12Giving.Grouped <- factor(cut(givings$FY12Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
givings$FY13Giving.Grouped <- factor(cut(givings$FY13Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
```

```r
givings$FY14Giving.Grouped <- factor(cut(givings$FY14Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
givings$FY15Giving.Grouped <- factor(cut(givings$FY15Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
givings$FY16Giving.Grouped <- factor(cut(givings$FY16Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
```

2.3 Univariate Data Analysis

We are going to conduct univariate data analysis for the following variables:

- Gender
- Class.Year
- Marital.Status
- Major
- Next.Degree
- AttendenceEvent
- FY12 though FY16 Giving (numerical, log transformed and Grouped)

2.3.1 Gender

```r
row <- xtabs(~Gender, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("Donor Count", "Ratio"))
```

```
##              Female    Male
## Donor Count 505.000 495.000
## Ratio         0.505   0.495
```

The dataset contains nearly identical amount of female vs. male donors. This result is mildly surprising but possible. According to the National Center for Education Statistics, the national average is 56% for female and 44% for male enrollment in college education (https://nces.ed.gov/programs/coe/indicator_cha.asp) in 2015. The data for graduation rates is similary skewed towards women. However, the rates are more likely to be skewed toward men in earlier years. Also, there is a chance that this specific university bucks the national trends for a variety of reasons.

2.3.2 Class.Year

```r
row <- xtabs(~Class.Year, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("Class.Year Count",
    "Ratio"))
```

```
##                    X1972   X1982   X1992   X2002   X2012
## Class.Year Count 105.000 176.000 203.000 223.000 293.000
## Ratio              0.105   0.176   0.203   0.223   0.293
```

This table is surprising. There are only 5 graduation years. The data is not a random subsample from the entire population but rather a subsample of 10-years (each data is 10 years apart). **It will be very difficult to argue that the results we infer from our model is applicable to all graduates of the university.**

2.3.4 Marital.Status

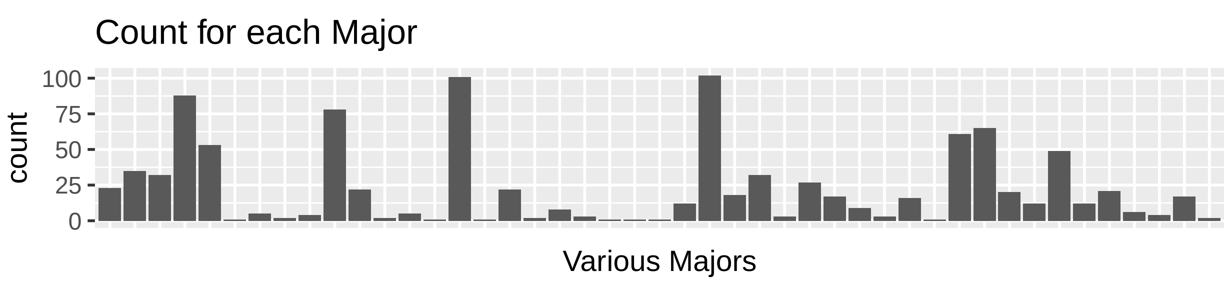
```
row <- xtabs(~Marital.Status, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("Marital.Status Count",
    "Ratio"))
```

```
##                      Divorced Married  Single Widowed
## Marital.Status Count   61.000 584.000 344.000  11.000
## Ratio                   0.061   0.584   0.344   0.011
```

Divorce to Marriage ratio is very low. According to Wikipedia (https://en.wikipedia.org/wiki/Divorce_demography), the expected ratio is around 44%. That said, the measurement methodology is slightly different and we expect rates to change with graduation years (divorce rates are more likely to increase with age). So we are going to assume that Marital.Status data is valid sample for the population.

2.3.5 Major

```
ggplot(givings, aes(x = Major)) + geom_histogram(stat = "count") + labs(title = "Count for each Major",
    x = "Various Majors") + theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



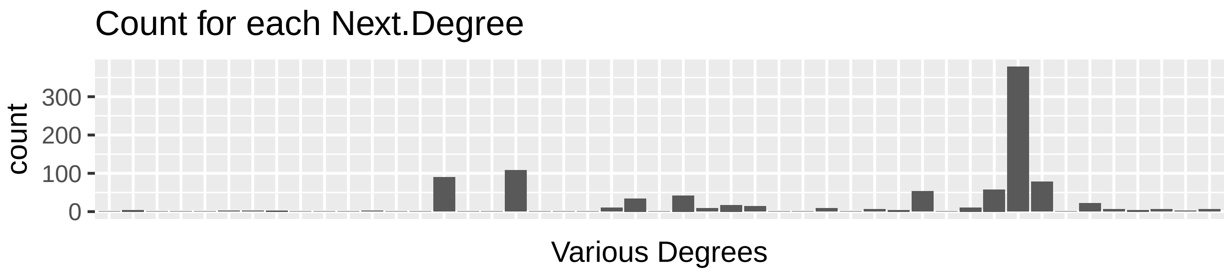

```
head(sort(xtabs(~Major, data = givings)), 3)
```

```
## Major
##            Chinese        Engineering English-Journalism
##                  1                  1                  1
```

Many of these factors have very little representation (ex: Chinese, English-Journalism) so we don't expect a significant contribution to our model. That said, we are going to investigate grouping strategies to improve our model.

2.3.6 Next.Degree

```
ggplot(givings, aes(x = Next.Degree)) + geom_histogram(stat = "count") + labs(title = "Count for each Ne
    x = "Various Degrees") + theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```




```r
givings$Adv.Deg <- fct_collapse(givings$Next.Degree, bachelor_equivalent = c("AA",
    "BA", "BAE", "BD", "BFA", "BN", "BS", "BSN", "LLB", "LLD", "NDA", "UBDS",
    "UDDS", "UMD", "UMDS", "UNKD", "TC"), above_bachelor = c("DC", "DDS", "DMD",
    "DO", "DO2", "DP", "JD", "PHD", "MA", "MA2", "MAE", "MALS", "MAT", "MBA",
    "MCP", "MD", "MD2", "ME", "MFA", "MHA", "ML", "MLS", "MM", "MPA", "MPH",
    "MS", "MSM", "MSW", "STM"))
```

The Next.Degree as a factor variable is too scathered. Many levels only have a single count (ex: MA2, MALS, MSM, BD, etc). We will group donor into 3 catories; those without a next degree (None), those with a bachelor equivalent and those with a degree higher than bachelor.

2.3.7 AttendenceEvent

```r
row <- xtabs(~AttendenceEvent, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("AttendenceEvent Count",
    "Ratio"))
```

```
##                         Didn.t.Attend Attended
## AttendenceEvent Count         395.000  605.000
## Ratio                           0.395    0.605
```

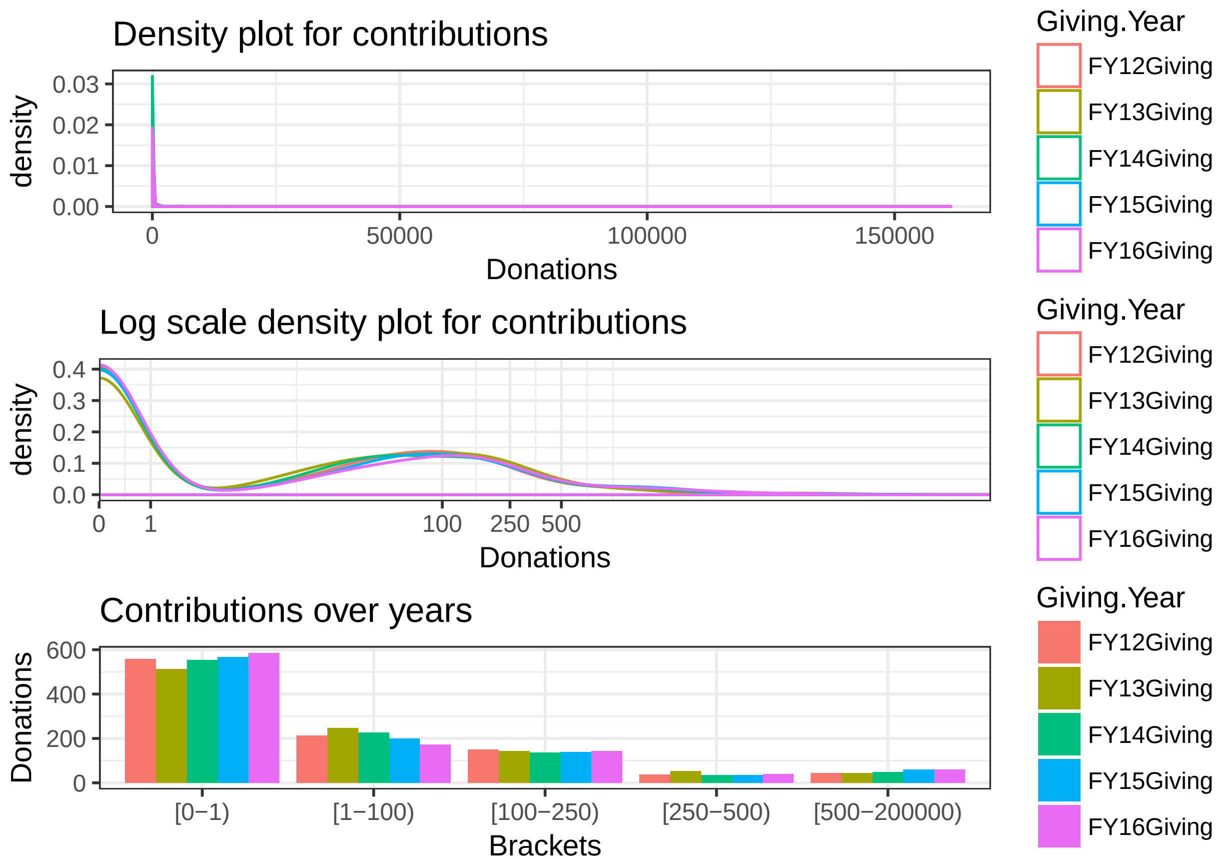
40% of graduates have attended at least one Alumni event organized between 2012 and 2015. This is a very high ratio. Intuitively, we expect a high correlation between this variable and donatios so we'll include this variable in our analysis.

2.3.8 FY12 though FY16 Giving (numerical, log transformed and Grouped)

```r
givings.tidy.donations <- givings[1:12] %>% gather("Giving.Year", "Donations",
    8:12)
givings.tidy.donations$Giving.Grouped <- factor(cut(givings.tidy.donations$Donations,
    breaks = c(0, 1, 100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)",
        "[250-500)", "[500-200000)"), right = FALSE))
givings.tidy.donations.aggregate <- as.data.frame(xtabs(~Giving.Grouped + Giving.Year,
    data = givings.tidy.donations))
p1 <- ggplot(givings.tidy.donations, aes(x = Donations, colour = Giving.Year)) +
    geom_density(alpha = 0.3) + labs(title = "Density plot for contributions") +
    theme_bw()

p2 <- ggplot(givings.tidy.donations, aes(x = Donations, colour = Giving.Year)) +
    geom_density(alpha = 0.3) + scale_x_continuous(breaks = c(0, 1, 100, 250,
    500, 2e+05), trans = "log1p", expand = c(0, 0)) + labs(title = "Log scale density plot for contribu
    scale_y_continuous() + theme_bw()
```

```
p3 <- ggplot(givings.tidy.donations.aggregate, aes(x = Giving.Grouped, y = Freq)) +
    geom_bar(aes(fill = Giving.Year), stat = "identity", position = "dodge") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) + labs(y = "Donations",
    x = "Brackets", title = "Contributions over years") + theme_bw()
grid.arrange(p1, p2, p3, ncol = 1, nrow = 3)
```



We note that continous scale values for the contributions have a very strong skew. At log scale, we have a bi-modal distribution, with most of the values centered around 0, and other around the $100 range. What may be important, however, is to uniquely indentify and model donors who are willing to make large contributions. We observe that most years follow a similar pattern. The donor behavior seems to be consistent over multiple years. We expect (and will confirm through bivariate analysis) to find a strong correlalation between a donor's 2016 preference and his/her previous years' preferences.

2.4 Bivariate Data Analysis

We are going to look at the relationship between these following variables:

- FY16Giving.Grouped vs. (Gender, Class.Year, Marital.Status, Major, Next.Degree, AttendenceEvent)
- FY16Giving.Grouped vs. (FY15Giving.Grouped, ..., FY12Giving.Grouped)
- Gender vs. (Class.Year, Marital.Status, Major)
- Major vs. Next.Degree
- Class.Year vs. AttendenceEvent

2.4.1 FY16Giving.Grouped vs. Gender

```
t1 <- xtabs(~Gender + FY16Giving.Grouped, data = givings)
t1.1 <- round(t1/rowSums(t1), 2)
# kable(list(t1, t1.1), caption = 'Frequency vs. Ratio for
# FY16Giving.Grouped vs. Gender')
t1
```

```
##         FY16Giving.Grouped
## Gender   [0-1) [1-100) [100-250) [250-500) [500-200000)
##   Female   298     106        58        17           26
##   Male     288      67        85        22           33
```

```
t1.1
```

```
##         FY16Giving.Grouped
## Gender   [0-1) [1-100) [100-250) [250-500) [500-200000)
##   Female  0.59    0.21      0.11      0.03         0.05
##   Male    0.58    0.14      0.17      0.04         0.07
```

In table 1, we note two interesting observations. There are more donations in the [$500-$200K) bracket than the [$250-$500) bracket. Also at $100 or above, men consistently donate more than women.

2.4.2 FY16Giving.Grouped vs. Class.Year

```
t2 <- xtabs(~Class.Year + FY16Giving.Grouped, data = givings)
t2.1 <- round(t2/rowSums(t2), 2)
# kable(list(t2,t2.1), caption = 'Frequency and Ratio for FY16Giving.Grouped
# vs. Class.Year')
t2
```

```
##           FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972    50       9        23         7           16
##       1982    90      22        35        14           15
##       1992   115      29        38         9           12
##       2002   137      41        25         6           14
##       2012   194      72        22         3            2
```

```
t2.1
```

```
##           FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972  0.48    0.09      0.22      0.07         0.15
##       1982  0.51    0.12      0.20      0.08         0.09
##       1992  0.57    0.14      0.19      0.04         0.06
##       2002  0.61    0.18      0.11      0.03         0.06
##       2012  0.66    0.25      0.08      0.01         0.01
```

There are 3 key insights from this table:

1. Older alumni make disportionately bigger donations (15% of the Class of 72 made $500 + donations).
2. A higher percentage of the older alumni make donations ($0 donations is only 48% for the class of 72, vs 66% for the class of 2012). 3- But there are more recent graduates (not sure why!). So even as their ratio is lower, most of the $1-$100 donations come from the class for 2012.

2.4.3 FY16Giving.Grouped vs. Marital.Status

```
t3 <- xtabs(~Marital.Status + FY16Giving.Grouped, data = givings)
t3.1 <- round(t3/rowSums(t3), 2)
# kable(list(t3,t3.1), caption = 'Frequency and Ratio for FY16Giving.Grouped
# vs. Marital.Status')
t3
```

```
##                 FY16Giving.Grouped
## Marital.Status [0-1) [1-100) [100-250) [250-500) [500-200000)
##       Divorced    36       9        11         2            3
##       Married    305      96       109        31           43
##       Single     241      66        23         4           10
##       Widowed      4       2         0         2            3
```

```
t3.1
```

```
##                 FY16Giving.Grouped
## Marital.Status [0-1) [1-100) [100-250) [250-500) [500-200000)
##       Divorced  0.59    0.15      0.18      0.03         0.05
##       Married   0.52    0.16      0.19      0.05         0.07
##       Single    0.70    0.19      0.07      0.01         0.03
##       Widowed   0.36    0.18      0.00      0.18         0.27
```

The data "appears" impressively clear. Married and single people are biggest source of donations. We expect Marital.Status to be a significant explanatory variable in our final model. That said, when we do the bi-variate analysis, we'll show that most of the data comes from from maried or single people, so the big skew in the ratio mostly attributable to the low number of divorced, widowed alumni in the data.

2.4.4 Donation level vs. Major

There are 45 majors in the dataset. Some majors only has one record. It is inappropriate to dump all majors as binary values into the model because:

1. It will cause curse of dimensionality and reduce prediction power.
2. These binary variables will holds most of records as zero and we barely have any infromation about them.

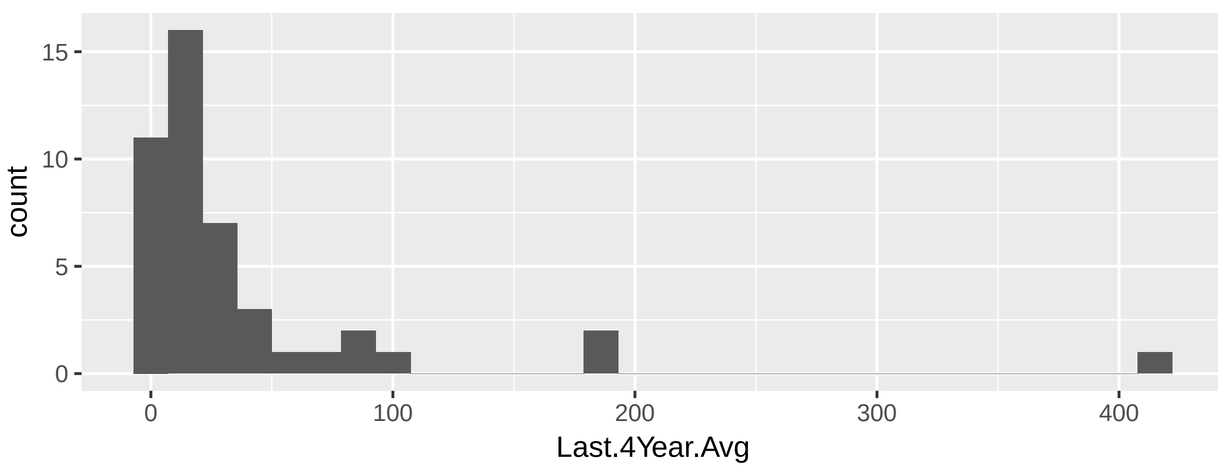
Therefore we need to group these majors. The method we used is to group major by the median donations in the last 4 years(2012-2015). First we calculate the last 4 year average donation for each person, then we check the median amount of last 4 year average for each major. Based on this value, we can label this major as somethings like "No","Low","Medium","High" donation major and put it into the model.

To find the right grouping way, we first start with a granular way to cut the median donation with 5 dollars increase. Please note that becuase it is a median value, it is much less extreme than the original donation amount(please check the histogram). The 50% of major median donation is less than 13.75 and only 25% of majors are larger than 33.75, so the original cut-offs [0, 1), [1, 100), [100, 250), [250, 500), [500, 200000) just won't work with because most of the values will be skewed in the [1:100). The granular way of cutting off increased the grouping by 5,like 0 0,1-6,6-11. We can observed the following patterns to get our final groupings:

1. There is an extreme value at 400 in the histogram. It came from the English Journalism. The English Jouralism only has one alumni in the sample and this person donate 1500 in 2015 and has donations every year from 2012 to 2015. This may indicates some problems in our sample because one person in a major can not represent the whole major. For now, **we may still label English Jouralism as a major with high donation level but will definitely need more data to justify this point if possible.**

2. The percentage of 2016 donation in [0,1) decreases as the it goes as major median donation goes higher. The percentage in higher buckets like [250-500) increases as the it goes as major median donation goes higher.
3. [0,1) is a natural cut-off point which means no one in this major donates
4. Majors with median donations in [1,10) are showing a similar behavior in 2016 donations (70% if alumni in these majors didn't donate in 2016). [10-35) are showing a similar behavior in 2016 donations, with 50-60% alumni no donations, 20% donating 1-100 and about 10% donating 100-250 and 5% in following two categories . Less than 25% of the major median are larger than 35, we can group them together.
5. We finally decide to go with the cut-off [0,1),[1,10),[10,35) and 35+ to group the group median donation values and give them No, low, medium and high labels. It is expected that the higher the major donation level is, the more donation the alumni makes in 2016. This is consistent with our third contingency table in this section. As the major donation_level increase from No to High, the percentage of 2016 donation in [0,1) decreases while the percentage in higher buckets like [250-500),[500-200000) increase.

```
givings$Last.4Year.Avg <- rowMeans(givings[c("FY12Giving", "FY13Giving", "FY14Giving",
    "FY15Giving")])
Major.Index <- data.frame(aggregate(Last.4Year.Avg ~ Major, data = givings,
    median))
ggplot(Major.Index, aes(Last.4Year.Avg)) + geom_histogram(bins = 30)
```




```
givings <- merge(x = givings, y = Major.Index, by = "Major")

givings$Major.Donation.Level <- factor(cut(givings$Last.4Year.Avg.y, labels = c("NO",
    "Low", "Medium", "High"), breaks = c(0, 1, 10, 30, 2e+05), right = FALSE))
t18 <- xtabs(~Major.Donation.Level + FY16Giving.Grouped, data = givings)
round(t18/rowSums(t18), 2)
```

```
##                     FY16Giving.Grouped
## Major.Donation.Level [0-1) [1-100) [100-250) [250-500) [500-200000)
##               NO      0.72    0.15      0.11      0.02         0.00
##               Low     0.67    0.16      0.08      0.06         0.03
##               Medium  0.58    0.18      0.15      0.03         0.07
##               High    0.48    0.13      0.22      0.10         0.08
```

```
givings$High.Donor.Major <- ifelse(givings$Major %in% c("History", "Psychology",
    "Biology", "Economics"), TRUE, FALSE)
```

We observe that Majors who (on median) made higher donations in previous years are more likely to make higher donations for FY2016.

2.4.5 FY16Giving.Grouped vs. Next.Degree

```
t12 <- xtabs(~Adv.Deg + FY16Giving.Grouped, data = givings)
round(t12/rowSums(t12), 2)
```

```
##                        FY16Giving.Grouped
## Adv.Deg                  [0-1) [1-100) [100-250) [250-500) [500-200000)
##   bachelor_equivalent  0.57    0.22      0.12      0.03          0.05
##   above_bachelor       0.49    0.21      0.17      0.04          0.09
##   NONE                 0.72    0.11      0.11      0.04          0.02
```

A higher proportion of people with above_bachelor degree make top donations ($500 or more), compared to other groups.

2.4.6 FY16Giving.Grouped vs. AttendenceEvent

```
(t4 <- xtabs(~AttendenceEvent + FY16Giving.Grouped, data = givings))
```

```
##                 FY16Giving.Grouped
## AttendenceEvent [0-1) [1-100) [100-250) [250-500) [500-200000)
##   Didn't Attend   286      61        36         5            7
##   Attended        300     112       107        34           52
```

The data is inline with our expectations. Among the people who donate, there is a strong correlation between attendence and donations. In fact, most of the top donors (52 out of 59, 85%) have attended an Alumni event.

2.4.7 FY16Giving.Grouped vs. previous years' Donation levels

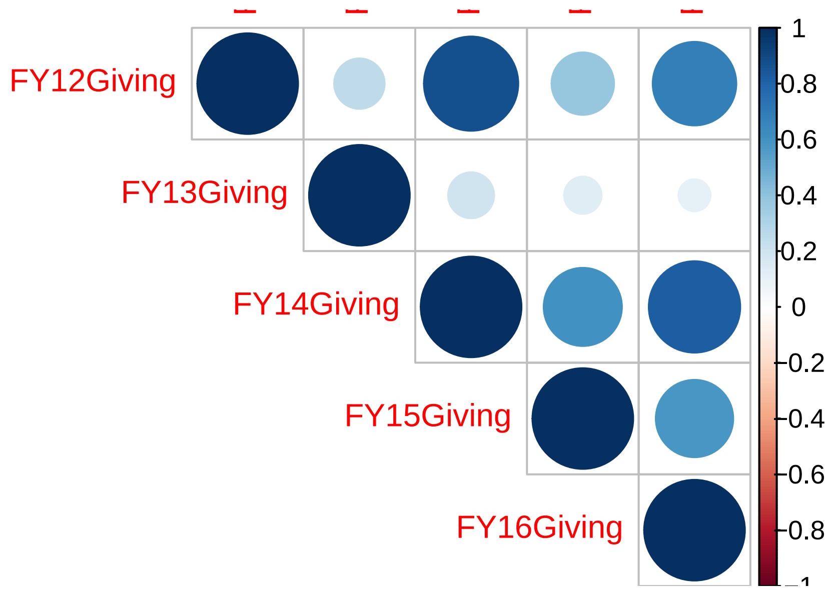
```
library(car)
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(corrplot)
M = givings[c(8:12)]
M_corr = cor(M)
corrplot(M_corr, method = "circle", type = "upper")
```




```
xtabs(~FY16Giving.Grouped + FY12Giving.Grouped, data = givings)
```

```
##                   FY12Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          462      73        38         9            4
##       [1-100)         60      96        16         0            1
##       [100-250)       26      40        69         5            3
##       [250-500)        4       2        16        16            1
##       [500-200000)     6       2        10         7           34
```

```
xtabs(~FY16Giving.Grouped + FY13Giving.Grouped, data = givings)
```

```
##                   FY13Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          441      94        40         7            4
##       [1-100)         39     123        10         1            0
##       [100-250)       19      27        73        19            5
##       [250-500)        5       0        13        18            3
##       [500-200000)     9       3         7         9           31
```

```
xtabs(~FY16Giving.Grouped + FY14Giving.Grouped, data = givings)
```

```
##                   FY14Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          461      82        34         4            5
##       [1-100)         56     108         8         1            0
##       [100-250)       23      33        74         8            5
##       [250-500)        5       2        15        17            0
##       [500-200000)     8       1         5         6           39
```

```
xtabs(~FY16Giving.Grouped + FY15Giving.Grouped, data = givings)
```

```
##                   FY15Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          480      64        29         5            8
##       [1-100)         57     108         8         0            0
##       [100-250)       23      25        88         4            3
##       [250-500)        3       1        10        23            2
```

```
##       [500-200000)      4        1          3          4            47
```

We notice that for any given donation bracket, most likely donation level for 2016 is the same level in 2015 (Ex: 40 out of 48 top donors in 2016 were also top donors in 2015). **The prior year donation levels are a strong indicator for this year's donation levels.** But, since the donation brackets are highly correletated, we expect FY15 to be strong indicator for FY16 and we also expect prior years' unique contributions to our model to be less impactful (due to the high colinearity).

2.4.8 Gender vs. Class.Year

```r
t6 <- xtabs(~Class.Year + Gender, data = givings)
t(t6)
```

```
##         Class.Year
## Gender   1972 1982 1992 2002 2012
##   Female   38   80  102  133  152
##   Male     67   96  101   90  141
```

```r
t(round(t6/rowSums(t6), 2))
```

```
##         Class.Year
## Gender   1972 1982 1992 2002 2012
##   Female 0.36 0.45 0.50 0.60 0.52
##   Male   0.64 0.55 0.50 0.40 0.48
```

As expected, over the years, the gender ratio converges towards a gender neutral 50%, but in the earlier years males make a higher percentage of the sample. It is also worth noting that there is an unexpected change in the ratio for the class of 2002. We will explore the Gender:Class.Year interaction

2.4.9 Gender vs. Marital.Status

```r
t7 <- xtabs(~Marital.Status + Gender, data = givings)
t(t7)
```

```
##         Marital.Status
## Gender   Divorced Married Single Widowed
##   Female       37     282    178       8
##   Male         24     302    166       3
```

```r
t(round(t7/rowSums(t7), 2))
```

```
##         Marital.Status
## Gender   Divorced Married Single Widowed
##   Female     0.61    0.48   0.52    0.73
##   Male       0.39    0.52   0.48    0.27
```

We knew that Married and Single people are more likely to donate. But now we understand that it is probably because vast majority of the data is made out of Married and Single people. We anticipate that this will weaking the predictive power of the Marital.Status variable. We also not that strong skew in widow ratio can be explained by the life expectency differences between men and women.

2.4.10 Gender vs. Major

We've already argued that, without a thoughtful grouping strategy, simply looking at each each Major is not insightful. We are going to explore the relationship between high/medium/low donation major groups and gender

```r
t8 <- xtabs(~Major.Donation.Level + Gender, data = givings)
t(t8)
```

```
##         Major.Donation.Level
## Gender    NO Low Medium High
##   Female  29  79    373   24
##   Male    25  38    364   68
```

```r
t(round(t8/rowSums(t8), 2))
```

```
##         Major.Donation.Level
## Gender     NO  Low Medium High
##   Female 0.54 0.68   0.51 0.26
##   Male   0.46 0.32   0.49 0.74
```

We had already established that among the high level donors, men had a higher ratio than women. We now conclude that this is also refected for Majors. Majors, that on average had lower previous year donations had a higher percentage of males vs females. And majors that have the highest donation levels have on more males than females.

2.4.11 Major vs. Next.Degree

skipping for now

2.4.12 Major vs. AttendenceEvent

```r
t9 <- xtabs(~Major.Donation.Level + AttendenceEvent, data = givings)
t(t9)
```

```
##                 Major.Donation.Level
## AttendenceEvent  NO Low Medium High
##   Didn't Attend  31  53    281   30
##   Attended       23  64    456   62
```

```r
t(round(t9/rowSums(t9), 2))
```

```
##                 Major.Donation.Level
## AttendenceEvent   NO  Low Medium High
##   Didn't Attend 0.57 0.45   0.38 0.33
##   Attended      0.43 0.55   0.62 0.67
```

Attendence level did have a positive impact on the donations. People from medium and high donations majors were more likely to have attended at least on event.

2.4.13 Class.Year vs. AttendenceEvent

```r
t10 <- xtabs(~Class.Year + AttendenceEvent, data = givings)
t(t10)
```

```
##                 Class.Year
## AttendenceEvent 1972 1982 1992 2002 2012
##   Didn't Attend   41   83   86   65  120
##   Attended        64   93  117  158  173
```

```r
t(round(t10/rowSums(t10), 2))
```

```
##                Class.Year
## AttendenceEvent 1972 1982 1992 2002 2012
##   Didn't Attend 0.39 0.47 0.42 0.29 0.41
##   Attended      0.61 0.53 0.58 0.71 0.59
```

It is remarkable that graduates from 1972 have the same (~60%) attendence rate (assuming the AttendenceEvent variable is for events since 2012 and not since they graduated) as the class of 2012. Class of 2002 has anusual spike, which is not explained by the available data.

2.5 Interactions

2.5.1 Gender, Class.Year, 2016 Donations

```r
t11 <- xtabs(~Class.Year + FY16Giving.Grouped + Gender, data = givings)
t11
```

```
## , , Gender = Female
##
##           FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972    21       3         4         4            6
##       1982    38      13        18         6            5
##       1992    61      17        15         3            6
##       2002    73      33        15         3            9
##       2012   105      40         6         1            0
##
## , , Gender = Male
##
##           FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972    29       6        19         3           10
##       1982    52       9        17         8           10
##       1992    54      12        23         6            6
##       2002    64       8        10         3            5
##       2012    89      32        16         2            2
```

The difference in top donations (above \$500) can be explained by male / female ratio. For example, 6 women made \$500+ donations from the class of 72, vs. 10 man. However, their ratio (6/10) is not too far from the female/male ratio (0.56) for the class of FY72. We are not anticipating a strong interaction between Gender, Class.Year, and 2016 Donation levels.

2.5.2 Gender, Major, 2016 donations

```r
t11 <- xtabs(~Major.Donation.Level + FY16Giving.Grouped + Gender, data = givings)
t11
```

```
## , , Gender = Female
##
##                     FY16Giving.Grouped
## Major.Donation.Level [0-1) [1-100) [100-250) [250-500) [500-200000)
##                   NO    23       5         1         0            0
```

```
##                 Low       54      15          2          4             4
##                 Medium   212      80         49         10            22
##                 High       9       6          6          3             0
##
## , , Gender = Male
##
##                        FY16Giving.Grouped
## Major.Donation.Level  [0-1) [1-100) [100-250) [250-500) [500-200000)
##                 NO        16       3          5          1             0
##                 Low       24       4          7          3             0
##                 Medium   213      54         59         12            26
##                 High      35       6         14          6             7
```

Looking at the top donors from Majors that have usually high donation levels, we see that they are more likely to be male. We know that top donor Majors (Major.Donation.Level == High) have a 3-to-1 Male/Female ratio. But we observe a 7-to-0 ration for Male/Female distribution for donors who have donated above $500 and are from top donor majors. **So we we believe there may be an interaction between Gender, Major, 2016 donations**

Talk to Kiesten about this stuff

There is a high correlation of the different fiscal years of giving with most other years (except fiscal year 2013). Maybe something went wrong with soliciting donations that year? NO - There was one very large donation by a former Science major that throws off correlations in 2013.

This is why it is good to use the grouped donation variable we are asked to create (done below).

Conclusion: Maybe should do some kind of grouping of years? But how - average continuous dollars for each year and then cut into groups? Or i guess we could just pick the most recent year (2014 for 2015 and 2015 for 2016). Below, there is a contingency table for categorical donation amount variables of 2015 and 2014 and it supports the idea of using the year before to model the current year.

There are usually less than 25 people every year who give 250-500 dollars or 500+ dollars. That could be problematic for accurately predicting who gives the highest donations.

3 Comparison between ordinal and nominal

```r
c1 <- xtabs(~Gender + FY16Giving.Grouped, data = givings)
c2 <- xtabs(~Marital.Status + FY16Giving.Grouped, data = givings)
c3 <- xtabs(~Class.Year + FY16Giving.Grouped, data = givings)
c4 <- xtabs(~Major.Donation.Level + FY16Giving.Grouped, data = givings)
c5 <- xtabs(~AttendenceEvent + FY16Giving.Grouped, data = givings)

odds_ratio <- function(r1) {
    df1 <- as.data.frame.matrix(r1)
    n <- dim(df1)[1]
    len <- dim(df1)[2]
    odds = data.frame(matrix(0, n, len - 1))
    colnames(odds) <- colnames(df1)[1:len - 1]
```

```r
    for (i in seq(1, len - 1)) {
        if (i == 1) {
            lowerp <- df1[, 1]
        } else {
            lowerp <- rowSums(df1[, 1:i])
        }
        if (i == len - 1) {
            upperp <- df1[, len]
        } else {
            upperp <- rowSums(df1[, (i + 1):len])
        }
        odds[, i] <- lowerp/upperp
    }
    round(odds, 2)

    oratio <- data.frame(matrix(0, n - 1, len - 1), row.names = rownames(df1)[2:n])
    colnames(oratio) <- colnames(odds)
    for (j in seq(1, n - 1)) oratio[j, ] <- odds[j + 1, ]/odds[j, ]
    return(round(oratio, 2))
}
# Gender
odds_ratio(c1)
```

```
##      [0-1) [1-100) [100-250) [250-500)
## Male  0.97    0.63      0.74      0.76
```

```r
# Marital Status
odds_ratio(c2)
```

```
##         [0-1) [1-100) [100-250) [250-500)
## Married  0.76    0.78      0.62      0.65
## Single   2.14    3.79      3.42      2.65
## Widowed  0.24    0.14      0.05      0.08
```

```r
# Graduating class
odds_ratio(c3)
```

```
##      [0-1) [1-100) [100-250) [250-500)
## 1982  1.15    1.36      1.42      1.93
## 1992  1.25    1.39      1.71      1.48
## 2002  1.22    1.62      1.17      0.94
## 2012  1.23    2.49      5.67      9.75
```

```r
# Major
odds_ratio(c4)
```

```
##        [0-1) [1-100) [100-250) [250-500)
## Low     0.77    0.72      0.18      0.00
## Medium  0.68    0.65      0.99      0.51
## High    0.67    0.50      0.50      0.85
```

```r
# Addent Event
odds_ratio(c5)
```

```
##          [0-1) [1-100) [100-250) [250-500)
## Attended  0.37     0.3      0.19      0.19
```

4 Modeling

```r
# i decided to estimate a model for FY15 categorical donations. Then we
# could use the model to evaluate how well we predicts FY16 donation
# patterns.

# i haven't looked at interaction terms yet! because i havent done bivariate
# analysis between explanatory variables yet.

# Proportional Odds Model don't forget to switch the sign of the
# coefficients from the polr function remember that if the coefficient for
# the overall variable is not significant, cannot use the coefficients for
# each category.(like in this case grouped_major)

library(ordinal)
```

```
##
## Attaching package: 'ordinal'

## The following object is masked from 'package:dplyr':
##
##     slice
```

```r
model_B1a <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped, data = givings,
    link = "logit")
summary(model_B1a)
```

```
## formula: FY16Giving.Grouped ~ FY15Giving.Grouped
## data:    givings
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  1000 -844.85 1705.71 6(0)  1.24e-09 9.0e+01
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## FY15Giving.Grouped[1-100)          2.0271     0.1715   11.82   <2e-16 ***
## FY15Giving.Grouped[100-250)        3.6406     0.2223   16.37   <2e-16 ***
## FY15Giving.Grouped[250-500)        5.5883     0.3755   14.88   <2e-16 ***
## FY15Giving.Grouped[500-200000)     7.6848     0.4388   17.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                          Estimate Std. Error z value
## [0-1)|[1-100)              1.6932     0.1164   14.54
## [1-100)|[100-250)          3.1028     0.1514   20.50
## [100-250)|[250-500)        5.3217     0.2341   22.74
## [250-500)|[500-200000)     6.6153     0.3067   21.57
```

```r
Anova(model_B1a)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    Df  Chisq Pr(>Chisq)
```

```
## FY15Giving.Grouped  4 628.74  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_B1b <- clm(formula = FY16Giving.Grouped ~ FY14Giving.Grouped, data = givings,
    link = "logit")
summary(model_B1b)
```

```
## formula: FY16Giving.Grouped ~ FY14Giving.Grouped
## data:    givings
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  1000 -901.09 1818.18 6(0)  8.32e-12 8.7e+01
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## FY14Giving.Grouped[1-100)          1.8216     0.1647   11.06   <2e-16 ***
## FY14Giving.Grouped[100-250)        3.2761     0.2154   15.21   <2e-16 ***
## FY14Giving.Grouped[250-500)        4.8554     0.3542   13.71   <2e-16 ***
## FY14Giving.Grouped[500-200000)     6.9691     0.4398   15.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                          Estimate Std. Error z value
## [0-1)|[1-100)              1.5905     0.1141   13.94
## [1-100)|[100-250)          2.8808     0.1436   20.06
## [100-250)|[250-500)        4.7857     0.2078   23.02
## [250-500)|[500-200000)     5.8204     0.2598   22.40
```

```
Anova(model_B1b)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    Df  Chisq Pr(>Chisq)
## FY14Giving.Grouped  4 625.66  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_B1c <- clm(formula = FY16Giving.Grouped ~ FY13Giving.Grouped, data = givings,
    link = "logit")
summary(model_B1c)
```

```
## formula: FY16Giving.Grouped ~ FY13Giving.Grouped
## data:    givings
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  1000 -908.39 1832.77 6(0)  9.03e-13 8.2e+01
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## FY13Giving.Grouped[1-100)          1.9178     0.1705   11.25   <2e-16 ***
## FY13Giving.Grouped[100-250)        3.3013     0.2180   15.15   <2e-16 ***
## FY13Giving.Grouped[250-500)        4.6617     0.3094   15.07   <2e-16 ***
## FY13Giving.Grouped[500-200000)     6.5887     0.4162   15.83   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                          Estimate Std. Error z value
## [0-1)|[1-100)              1.7805     0.1272   14.00
## [1-100)|[100-250)          3.0812     0.1550   19.87
## [100-250)|[250-500)        4.9061     0.2105   23.30
## [250-500)|[500-200000)     5.7924     0.2480   23.36
```

```r
Anova(model_B1c)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                      Df  Chisq Pr(>Chisq)
## FY13Giving.Grouped  4 628.55  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_B1d <- clm(formula = FY16Giving.Grouped ~ FY12Giving.Grouped, data = givings,
    link = "logit")
summary(model_B1d)
```

```
## formula: FY16Giving.Grouped ~ FY12Giving.Grouped
## data:    givings
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  1000 -928.44 1872.87 6(0)  1.16e-13 8.9e+01
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## FY12Giving.Grouped[1-100)          1.8768     0.1656   11.34   <2e-16 ***
## FY12Giving.Grouped[100-250)        3.0601     0.2034   15.04   <2e-16 ***
## FY12Giving.Grouped[250-500)        4.3496     0.3589   12.12   <2e-16 ***
## FY12Giving.Grouped[500-200000)     6.6054     0.4455   14.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                          Estimate Std. Error z value
## [0-1)|[1-100)              1.5607     0.1120   13.94
## [1-100)|[100-250)          2.7763     0.1383   20.08
## [100-250)|[250-500)        4.5059     0.1924   23.41
## [250-500)|[500-200000)     5.4599     0.2397   22.77
```

```r
Anova(model_B1d)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                      Df  Chisq Pr(>Chisq)
## FY12Giving.Grouped  4 638.66  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_B2 <- clm(formula = FY16Giving.Grouped ~ Adv.Deg, data = givings, link = "logit")
summary(model_B2)
```

```
## formula: FY16Giving.Grouped ~ Adv.Deg
## data:    givings
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  1000 -1165.13 2342.25 7(0)  2.47e-10 1.3e+02
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## Adv.Degabove_bachelor   0.3746     0.1929   1.942  0.05219 .
## Adv.DegNONE            -0.5628     0.2077  -2.710  0.00673 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                         Estimate Std. Error z value
## [0-1)|[1-100)             0.3472     0.1750   1.984
## [1-100)|[100-250)         1.1763     0.1791   6.567
## [100-250)|[250-500)       2.2649     0.1957  11.574
## [250-500)|[500-200000)   2.8200     0.2124  13.279
```

```r
Anova(model_B2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##         Df  Chisq Pr(>Chisq)
## Adv.Deg  2 210.13  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_B3a <- clm(formula = FY16Giving.Grouped ~ Major.Donation.Level + FY15Giving.Grouped,
    data = givings, link = "logit")
summary(model_B3a)
```

```
## formula: FY16Giving.Grouped ~ Major.Donation.Level + FY15Giving.Grouped
## data:    givings
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  1000 -844.04 1710.07 6(0)  1.26e-09 2.3e+02
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## Major.Donation.LevelLow        -0.01957    0.38437  -0.051    0.959
## Major.Donation.LevelMedium      0.18764    0.33048   0.568    0.570
## Major.Donation.LevelHigh        0.32890    0.39059   0.842    0.400
## FY15Giving.Grouped[1-100)       2.02608    0.17179  11.794   <2e-16 ***
## FY15Giving.Grouped[100-250)     3.61527    0.22311  16.204   <2e-16 ***
## FY15Giving.Grouped[250-500)     5.56162    0.37704  14.751   <2e-16 ***
## FY15Giving.Grouped[500-200000)  7.66703    0.43952  17.444   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Threshold coefficients:
##                        Estimate Std. Error z value
## [0-1)|[1-100)            1.8534     0.3289   5.635
## [1-100)|[100-250)        3.2650     0.3426   9.529
## [100-250)|[250-500)      5.4853     0.3865  14.194
## [250-500)|[500-200000)   6.7794     0.4346  15.598
```

```
Anova(model_B3a)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                       Df  Chisq Pr(>Chisq)
## Major.Donation.Level  3 388.42  < 2.2e-16 ***
## FY15Giving.Grouped    4 465.78  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_B3b <- clm(formula = FY16Giving.Grouped ~ High.Donor.Major + FY15Giving.Grouped,
    data = givings, link = "logit")
summary(model_B1b)
```

```
## formula: FY16Giving.Grouped ~ FY14Giving.Grouped
## data:    givings
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  1000 -901.09 1818.18 6(0)  8.32e-12 8.7e+01
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## FY14Giving.Grouped[1-100)        1.8216     0.1647   11.06   <2e-16 ***
## FY14Giving.Grouped[100-250)      3.2761     0.2154   15.21   <2e-16 ***
## FY14Giving.Grouped[250-500)      4.8554     0.3542   13.71   <2e-16 ***
## FY14Giving.Grouped[500-200000)   6.9691     0.4398   15.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                        Estimate Std. Error z value
## [0-1)|[1-100)            1.5905     0.1141   13.94
## [1-100)|[100-250)        2.8808     0.1436   20.06
## [100-250)|[250-500)      4.7857     0.2078   23.02
## [250-500)|[500-200000)   5.8204     0.2598   22.40
```

```
Anova(model_B3b)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                     Df  Chisq Pr(>Chisq)
## High.Donor.Major    1 178.83  < 2.2e-16 ***
## FY15Giving.Grouped  4 626.85  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_B3a, model_B3b)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_B3b FY16Giving.Grouped ~ High.Donor.Major + FY15Giving.Grouped
## model_B3a FY16Giving.Grouped ~ Major.Donation.Level + FY15Giving.Grouped
##           link: threshold:
## model_B3b logit flexible
## model_B3a logit flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B3b      9 1707.7 -844.83
## model_B3a     11 1710.1 -844.04  1.5849  2     0.4527
```

it seems model1c,Class.Year +Marital.Status + AttendenceEvent+FY15Giving.Grouped works best for now. The major variable is neither significant or contributing to the model. The FY15Giving works better than putting all past years The FY14Giving works better than using last 4 year average

```r
# they kind of suggested we should compare against multinomial regression
library(nnet)

model2 <- multinom(formula = FY15Giving.Grouped ~ Class.Year + Major.Donation.Level +
    Marital.Status + AttendenceEvent + FY14Giving.Grouped, data = givings)
```

```
## # weights:  85 (64 variable)
## initial  value 1609.437912
## iter  10 value 773.662374
## iter  20 value 729.551520
## iter  30 value 724.763633
## iter  40 value 723.733132
## iter  50 value 723.479313
## iter  60 value 723.465060
## final  value 723.464803
## converged
```

```r
Anova(model2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY15Giving.Grouped
##                      LR Chisq Df Pr(>Chisq)
## Class.Year              38.64 16   0.001225 **
## Major.Donation.Level    12.22 12   0.428271
## Marital.Status          11.53 12   0.483989
## AttendenceEvent         17.86  4   0.001315 **
## FY14Giving.Grouped     689.66 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model2a <- multinom(formula = FY15Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + FY14Giving.Grouped, data = givings)
```

```
## # weights:  70 (52 variable)
## initial  value 1609.437912
## iter  10 value 804.303016
## iter  20 value 734.569065
```

```
## iter  30 value 730.473759
## iter  40 value 729.697717
## iter  50 value 729.580091
## iter  60 value 729.574229
## final  value 729.574204
## converged
```

```r
Anova(model2a)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY15Giving.Grouped
##                    LR Chisq Df Pr(>Chisq)
## Class.Year            42.58 16  0.0003231 ***
## Marital.Status        13.81 12  0.3128510
## AttendenceEvent       19.08  4  0.0007597 ***
## FY14Giving.Grouped   694.27 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# use all past years
model2b <- multinom(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + FY12Giving + FY13Giving + FY14Giving + FY15Giving, data = givings)
```

```
## # weights:  70 (52 variable)
## initial  value 1609.437912
## iter  10 value 1562.571980
## iter  20 value 1280.186798
## iter  30 value 1045.313923
## iter  40 value 965.825083
## iter  50 value 954.435606
## iter  60 value 945.912119
## iter  70 value 943.913792
## iter  80 value 943.784608
## iter  90 value 943.763991
## iter 100 value 943.763790
## final  value 943.763790
## stopped after 100 iterations
```

```r
summary(model2b)
```

```
## Call:
## multinom(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
##     AttendenceEvent + FY12Giving + FY13Giving + FY14Giving +
##     FY15Giving, data = givings)
##
## Coefficients:
##              (Intercept) Class.Year1982 Class.Year1992 Class.Year2002
## [1-100)        -1.956978      0.3289484      0.3342477      0.4798835
## [100-250)      -2.204312     -0.1600977     -0.3173259     -0.8770671
## [250-500)      -5.484419      0.5647529     -0.2330442     -0.8746207
## [500-200000)   -5.680228      0.5176691     -0.6690780      0.2307101
##              Class.Year2012 Marital.StatusMarried Marital.StatusSingle
## [1-100)           0.8317810            0.04673449          -0.29924204
## [100-250)        -0.8003570            0.41988301          -0.35218326
## [250-500)        -0.8887495            0.97207842          -0.07627183
```

```
## [500-200000)     -0.7442782           0.17464936           0.19400098
##               Marital.StatusWidowed AttendenceEventAttended  FY12Giving
## [1-100)                   0.7364573               0.5322993 0.001807188
## [100-250)                -6.8528319               0.9694694 0.002511191
## [250-500)                 2.4710234               2.0397193 0.006182405
## [500-200000)              2.0172297               1.8189197 0.007524561
##                  FY13Giving  FY14Giving    FY15Giving
## [1-100)      -0.0025405533 0.001847834 -0.0010453033
## [100-250)    -0.0001868833 0.005204773  0.0006795244
## [250-500)    -0.0013988960 0.003354256  0.0023992496
## [500-200000) -0.0027172630 0.005065187  0.0031523546
##
## Std. Errors:
##              (Intercept) Class.Year1982 Class.Year1992 Class.Year2002
## [1-100)       0.09270662      0.1978350      0.1795664     0.15944685
## [100-250)     0.11662325      0.1844851      0.1826351     0.19643923
## [250-500)     0.09250048      0.2279492      0.1237746     0.07694984
## [500-200000)  0.10843300      0.1437613      0.0347978     0.11217965
##              Class.Year2012 Marital.StatusMarried Marital.StatusSingle
## [1-100)          0.14330138             0.1083858            0.1082423
## [100-250)        0.19855546             0.1363143            0.1496865
## [250-500)        0.03699398             0.1437273            0.1129554
## [500-200000)     0.03769735             0.2131544            0.2003913
##              Marital.StatusWidowed AttendenceEventAttended  FY12Giving
## [1-100)               7.936287e-03              0.18334770 0.001798099
## [100-250)             7.131247e-06              0.22143666 0.001238326
## [250-500)             2.962337e-02              0.09959312 0.001513339
## [500-200000)          2.226205e-02              0.10789582 0.001526949
##                FY13Giving  FY14Giving   FY15Giving
## [1-100)      0.0017224851 0.002063749 0.0015293378
## [100-250)    0.0000594158 0.001291722 0.0007669085
## [250-500)    0.0008420827 0.001492180 0.0007633876
## [500-200000) 0.0008974927 0.001402633 0.0007222133
##
## Residual Deviance: 1887.528
## AIC: 1991.528
```

```r
Anova(model2b)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                 LR Chisq Df Pr(>Chisq)
## Class.Year        30.724 16    0.01459 *
## Marital.Status    22.889 12    0.02868 *
## AttendenceEvent   41.330  4  2.296e-08 ***
## FY12Giving        27.706  4  1.430e-05 ***
## FY13Giving        60.590  4  2.180e-12 ***
## FY14Giving        23.681  4  9.256e-05 ***
## FY15Giving        29.691  4  5.656e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# use major values as index
model2c <- multinom(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
```

```r
    AttendenceEvent + FY14Giving.Grouped + Last.4Year.Avg.y + FY13Giving.Grouped +
    FY12Giving.Grouped + FY15Giving.Grouped, data = givings)
```

```
## # weights:  135 (104 variable)
## initial  value 1609.437912
## iter  10 value 892.448560
## iter  20 value 613.460425
## iter  30 value 594.506345
## iter  40 value 592.603025
## iter  50 value 592.215179
## iter  60 value 592.012274
## iter  70 value 591.890154
## iter  80 value 591.867171
## iter  90 value 591.866028
## final  value 591.866015
## converged
```

```r
Anova(model2c)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    LR Chisq Df Pr(>Chisq)
## Class.Year           12.540 16  0.7060477
## Marital.Status       12.178 12  0.4314761
## AttendenceEvent       8.745  4  0.0678097 .
## FY14Giving.Grouped   31.138 16  0.0129213 *
## Last.4Year.Avg.y      5.841  4  0.2113242
## FY13Giving.Grouped   73.164 16  2.768e-09 ***
## FY12Giving.Grouped   40.313 16  0.0007006 ***
## FY15Giving.Grouped  190.923 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# use major values as a category
model2d <- multinom(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level, data = givings)
```

```
## # weights:  85 (64 variable)
## initial  value 1609.437912
## iter  10 value 790.017919
## iter  20 value 722.479621
## iter  30 value 715.518408
## iter  40 value 714.318526
## iter  50 value 714.114177
## iter  60 value 714.020116
## iter  70 value 714.018885
## final  value 714.018870
## converged
```

```r
Anova(model2d)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                      LR Chisq Df Pr(>Chisq)
```

```
## Class.Year              17.32 16  0.3648974
## Marital.Status          17.20 12  0.1422784
## AttendenceEvent         18.80  4  0.0008599 ***
## FY15Giving.Grouped     715.93 16  < 2.2e-16 ***
## Major.Donation.Level    15.77 12  0.2022193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# only use 2015
model2e <- multinom(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + FY15Giving.Grouped, data = givings)
```

```
## # weights:  70 (52 variable)
## initial  value 1609.437912
## iter  10 value 776.849043
## iter  20 value 727.211432
## iter  30 value 722.338838
## iter  40 value 721.932102
## iter  50 value 721.910177
## iter  60 value 721.902908
## iter  70 value 721.902219
## final  value 721.901568
## converged
```

```r
Anova(model2e)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    LR Chisq Df Pr(>Chisq)
## Class.Year            17.43 16  0.3580986
## Marital.Status        16.44 12  0.1719384
## AttendenceEvent       18.82  4  0.0008525 ***
## FY15Giving.Grouped   721.48 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Use last 4 year avg(no major variable)
model2f <- multinom(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + Last.4Year.Avg.x, data = givings)
```

```
## # weights:  55 (40 variable)
## initial  value 1609.437912
## iter  10 value 1175.343396
## iter  20 value 1078.674332
## iter  30 value 1064.474374
## iter  40 value 1061.741486
## iter  50 value 1061.552345
## iter  60 value 1061.544935
## final  value 1061.544807
## converged
```

```r
Anova(model2f)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
```

```
##                   LR Chisq Df Pr(>Chisq)
## Class.Year          58.219 16  1.042e-06 ***
## Marital.Status      30.599 12   0.002267 **
## AttendenceEvent     71.738  4  9.750e-15 ***
## Last.4Year.Avg.x    42.196  4  1.519e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# comparison between different models
anova(model2e, model2b)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: FY16Giving.Grouped
##                                                                                         Model
## 1                                   Class.Year + Marital.Status + AttendenceEvent + FY15Giving.Grouped
## 2 Class.Year + Marital.Status + AttendenceEvent + FY12Giving + FY13Giving + FY14Giving + FY15Giving
##   Resid. df Resid. Dev   Test    Df  LR stat. Pr(Chi)
## 1      3948   1443.803
## 2      3948   1887.528 1 vs 2     0 -443.7244       1
```

```r
anova(model2e, model2c)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: FY16Giving.Grouped
##
## 1                                                                                Class.Year + Mar
## 2 Class.Year + Marital.Status + AttendenceEvent + FY14Giving.Grouped + Last.4Year.Avg.y + FY13Giving
##   Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1      3948   1443.803
## 2      3896   1183.732 1 vs 2    52 260.0711       0
```

```r
anova(model2e, model2d)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: FY16Giving.Grouped
##                                                                                  Model
## 1                         Class.Year + Marital.Status + AttendenceEvent + FY15Giving.Grouped
## 2 Class.Year + Marital.Status + AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level
##   Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1      3948   1443.803
## 2      3936   1428.038 1 vs 2    12 15.76539 0.2022193
```

```r
anova(model2e, model2f)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: FY16Giving.Grouped
##                                                                  Model
## 1   Class.Year + Marital.Status + AttendenceEvent + Last.4Year.Avg.x
## 2 Class.Year + Marital.Status + AttendenceEvent + FY15Giving.Grouped
##   Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1      3960   2123.090
## 2      3948   1443.803 1 vs 2    12 679.2865       0
```

it seems model2e,Class.Year +Marital.Status + AttendenceEvent+FY15Giving.Grouped works best for now.

The major variable is neither significant or contributing to the model. The FY15Giving works better than putting all past years The FY14Giving works better than using last 4 year average

We might want to switch base level of Class.Year to 2012 instead of 1972. As EDA suggests, Major doens't seem significant in either model.

Predictive problems:

-very few of some of majors, not good ability to predict how much people with those majors are likely to donate - this may not matter depending on whether major is significant in our final model.

Also, there very few large donations every year -usually less than 25 people who give 250-500 dollars and less than 25 people who give 500+ dollars. It may be hard to predict who are the highest donors. For instance, it seems like men may make large donations more frequently than women, but the difference is not significantly different. Maybe if we had higher numbers, we would see a trend exists there?

I will create a "next degree"" grouping (i didn't explore this variable yet): thinking of doing "bachelors", "graduate level degree", "none". Or should i make it more granular ie. split into MBA, MD, JD, etc ? Or try to guess which degrees are "professional degrees" - like MBA, MD, JD vs masters and PhD?