W271 Section 3 Lab 2

Kiersten Henderson, Jill Zhang, Hoang Phan, Daghan Altas

10/8/2017

```
#knitr::opts_chunk$set(cache=TRUE)
library(knitr)
library(vcd)
```

```
## Loading required package: grid
```

```
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

1 Introduction

An introduction to the project, which should include a concise summary of the key results as well as techniques you used in your final model.

2 Exploratory Data Analysis

```
library(Hmisc)
```

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

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:Hmisc':
##
##     combine, src, summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

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

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## between():   dplyr, data.table
## combine():   dplyr, Hmisc
## filter():    dplyr, stats
## first():     dplyr, data.table
## lag():       dplyr, stats
## last():      dplyr, data.table
## src():       dplyr, Hmisc
## summarize(): dplyr, Hmisc
## transpose(): purrr, data.table
```

```
library(forcats)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
## The following object is masked from 'package:Hmisc':
##
##     combine
```

```
if (dir.exists("XXXX")) {
    ## For Kiersten
    setwd("XXXX")
} else if (dir.exists("YYY")) {
    ## For Jill
    setwd("YYY")
} else if (dir.exists("ZZZ")) {
    ## For Hoang
    setwd("ZZZ")
} else if (dir.exists("/Users/daghan/Hacking/Berkeley/W271/Labs/w271_lab2/")) {
    ## For Daghan home computer
    setwd("/Users/daghan/Hacking/Berkeley/W271/Labs/w271_lab2/")
}

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

```
## givings
##
##  12  Variables      1000  Observations
## ---------------------------------------------------------------------------
## X
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0     1000        1    615.4    410.6    62.95   122.90
##      .25      .50      .75      .90      .95
##   308.75   613.00   917.25  1110.30  1174.05
##
## lowest :    1    2    3    4    5, highest: 1225 1226 1228 1229 1230
## ---------------------------------------------------------------------------
## Gender
##        n  missing distinct
##     1000        0        2
##
## Value          F     M
## Frequency    505   495
## Proportion 0.505 0.495
## ---------------------------------------------------------------------------
## Class.Year
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        5    0.949     1996    15.07
##
## Value       1972  1982  1992  2002  2012
```

```
## Frequency    105   176   203   223   293
## Proportion 0.105 0.176 0.203 0.223 0.293
## ---------------------------------------------------------------------------
## Marital.Status
##        n  missing distinct
##     1000        0        4
##
## Value          D     M     S     W
## Frequency     61   584   344    11
## Proportion 0.061 0.584 0.344 0.011
## ---------------------------------------------------------------------------
## Major
##        n  missing distinct
##     1000        0       45
##
## lowest : American Studies     Anthropology         Art                  Biology              Chemist
## highest: Spanish              Speech (Drama, etc.) Speech Correction    Theatre              Zoology
## ---------------------------------------------------------------------------
## Next.Degree
##        n  missing distinct
##     1000        0       47
##
## lowest : AA   BA   BAE  BD   BFA , highest: UBDS UDDS UMD  UMDS UNKD
## ---------------------------------------------------------------------------
## AttendenceEvent
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1000        0        2    0.717      605    0.605   0.4784
##
## ---------------------------------------------------------------------------
## FY12Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       66    0.826    186.9    345.5        0        0
##      .25      .50      .75      .90      .95
##        0        0       60      200      350
##
## lowest :     0.00     5.00     6.50     7.00     8.00
## highest: 10000.00 12000.00 16959.99 20000.00 21000.00
## ---------------------------------------------------------------------------
## FY13Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       78    0.864    311.5    590.4      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.0     75.0    210.5    400.0
##
## Value           0    500   1000   1500   2000   2500   3000   5000   5500
## Frequency     920     48     13      4      2      3      2      2      1
## Proportion  0.920  0.048  0.013  0.004  0.002  0.003  0.002  0.002  0.001
##
## Value        8000  12000  13000  14500 161500
## Frequency       1      1      1      1      1
## Proportion  0.001  0.001  0.001  0.001  0.001
## ---------------------------------------------------------------------------
## FY14Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
```

```
##     1000        0       80     0.83    142.6    255.5        0        0
##      .25      .50      .75      .90      .95
##        0        0       50      200      450
##
## lowest :     0.00     1.00     5.00     8.00    10.00
## highest:  5000.00  6000.00  8031.00 10000.00 11187.26
## ------------------------------------------------------------------------
## FY15Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       62    0.817    252.2    470.7      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.0     75.0    200.0    538.3
##
## lowest :     0.0     5.0    10.0    13.0    15.0
## highest: 10000.0 14776.0 15634.5 26500.0 58785.5
## ------------------------------------------------------------------------
## FY16Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       71    0.798      170    308.2        0        0
##      .25      .50      .75      .90      .95
##        0        0       75      216      500
##
## lowest :     0.00     5.00    10.00    15.00    18.00
## highest:  5000.00  6500.00 11500.00 11505.84 14655.25
## ------------------------------------------------------------------------
```

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)
```

We are also going to group majors into main categories including Sciences, SocialScience, FineArts, ForeignLanguages, Education, PhilosophyReligion, Mathematics, MultidisciplinaryStudies, English_Literature, Business_Economics, and ComputerScience_Engineering.

```
givings$Grouped.Major <- fct_collapse(givings$Major, Sciences = c("Chemistry",
    "Physics", "Biology", "General Science-Chemistry", "General Science-Psycho",
    "General Science-Math", "Mathematics-Physics", "General Science-Biology",
    "Zoology", "General Science", "General Science-Physics"),
    SocialScience = c("Psychology", "Anthropology", "Sociology",
        "History", "Political Science", "Sociology-Anthropology",
        "Economics-Regional Stds.", "Pol. Sci.-Regional Stds."),
    FineArts = c("Theatre", "Music", "Art", "Speech (Drama, etc.)"),
    ForeignLanguages = c("Spanish", "German", "French", "Russian",
        "Chinese"), Education = c("Physical Education", "Education",
        "Speech Correction"), PhilosophyReligion = c("Religious Studies",
        "Philosophy", "Philosophy-Religion"), Mathematics = "Mathematics",
    MultidisciplinaryStudies = "Independent", English_Literature = c("English",
        "Classics", "American Studies", "Comparative Literature",
        "English-Journalism"), Business_Economics = c("Economics-Business",
        "Economics"), ComputerScience_Engineering = c("Computer Science",
        "Engineering"))
```

Lastly, 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)"),
    include.lowest = TRUE))
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)"),
    include.lowest = TRUE))
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)"),
    include.lowest = TRUE))
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)"),
    include.lowest = TRUE))
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)"),
    include.lowest = TRUE))
```

2.3 Univariate Data Analysis

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

- Gender
- Class.Year
- Marital.Status
- Major and Grouped.Major
- Next.Degree

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

2.3.1 Gender

```
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

```
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 and Grouped.Major

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

```
##             American.Studies Anthropology    Art Biology Chemistry Chinese
## Major Count           23.000       35.000 32.000  88.000    53.000   1.000
## Ratio                  0.023        0.035  0.032   0.088     0.053   0.001
##             Classics Comparative.Literature Computer.Science Economics
## Major Count    5.000                 2.000            4.000    78.000
## Ratio          0.005                 0.002            0.004     0.078
##             Economics.Business Economics.Regional.Stds. Education
## Major Count             22.000                    2.000     5.000
## Ratio                    0.022                    0.002     0.005
##             Engineering English English.Journalism French General.Science
## Major Count       1.000 101.000             1.000 22.000           2.000
## Ratio             0.001   0.101             0.001  0.022           0.002
##             General.Science.Biology General.Science.Chemistry
## Major Count                   8.000                     3.000
## Ratio                         0.008                     0.003
##             General.Science.Math General.Science.Physics
## Major Count                1.000                   1.000
## Ratio                      0.001                   0.001
##             General.Science.Psycho German History Independent Mathematics
## Major Count                  1.000 12.000 102.000      18.000      32.000
## Ratio                        0.001  0.012   0.102       0.018       0.032
##             Mathematics.Physics  Music Philosophy Philosophy.Religion
## Major Count               3.000 27.000     17.000               9.000
## Ratio                     0.003  0.027      0.017               0.009
##             Physical.Education Physics Pol..Sci..Regional.Stds.
## Major Count              3.000  16.000                   1.000
## Ratio                    0.003   0.016                   0.001
##             Political.Science Psychology Religious.Studies Russian
## Major Count            61.000     65.000             20.00  12.000
## Ratio                   0.061      0.065              0.02   0.012
##             Sociology Sociology.Anthropology Spanish Speech..Drama..etc..
## Major Count    49.000                12.000  21.000               6.000
## Ratio           0.049                 0.012   0.021               0.006
##             Speech.Correction Theatre Zoology
## Major Count             4.000  17.000   2.000
## Ratio                   0.004   0.017   0.002
```

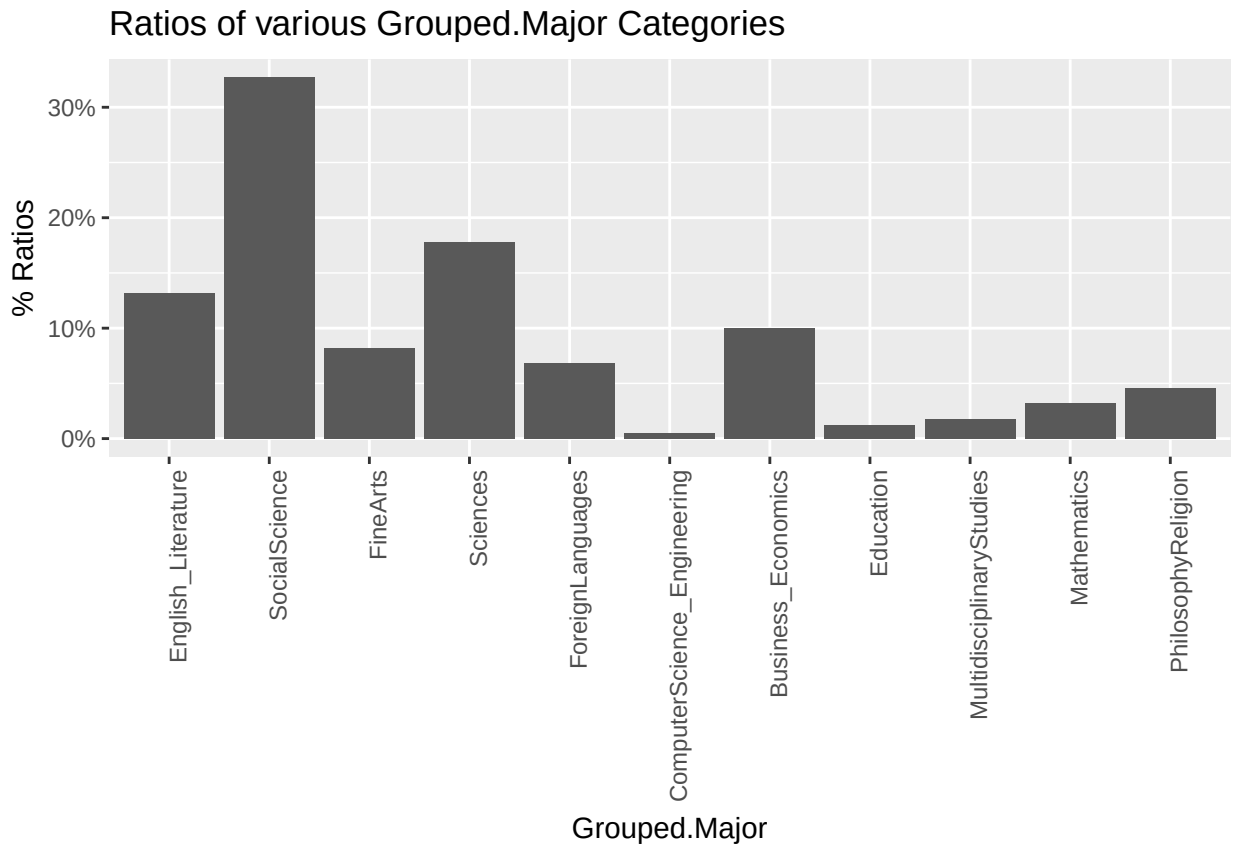
Many of these factors have very little representation (ex: Zoology, Political stuies in regional studies) so it makes sense to group them into major categories

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

```
##                     English_Literature SocialScience FineArts Sciences
## Grouped.Major Count            132.000       327.000   82.000  178.000
## Ratio                            0.132         0.327    0.082    0.178
##                     ForeignLanguages ComputerScience_Engineering
## Grouped.Major Count           68.000                       5.000
```

```
## Ratio                          0.068                       0.005
##                      Business_Economics Education MultidisciplinaryStudies
## Grouped.Major Count              100.0    12.000                    18.000
## Ratio                              0.1     0.012                     0.018
##                      Mathematics PhilosophyReligion
## Grouped.Major Count       32.000             46.000
## Ratio                      0.032              0.046
```

```
ggplot(givings, aes(x = Grouped.Major)) + geom_bar(aes(y = (..count..)/sum(..count..))) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    scale_y_continuous(labels = percent) + labs(y = "% Ratios",
    title = "Ratios of various Grouped.Major Categories")
```



2.3.6 Next.Degree

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

```
##                       AA    BA   BAE    BD   BFA    BN    BS   BSN    DC
## Next.Degree Count 1.000 4.000 1.000 1.000 1.000 2.000 2.000 3.000 1.000
## Ratio             0.001 0.004 0.001 0.001 0.001 0.002 0.002 0.003 0.001
##                     DDS   DMD    DO   DO2    DP    JD   LLB   LLD      MA
## Next.Degree Count 1.000 1.000 2.000 1.000 1.000 90.00 1.000 1.000 108.000
## Ratio             0.001 0.001 0.002 0.001 0.001  0.09 0.001 0.001   0.108
##                     MA2   MAE  MALS   MAT    MBA   MCP     MD   MD2     ME
```

```
## Next.Degree Count 1.000 1.000 1.000 10.00 34.000 1.000 42.000 9.000 17.000
## Ratio             0.001 0.001 0.001  0.01  0.034 0.001  0.042 0.009  0.017
##                      MFA   MHA    ML   MLS    MM   MPA   MPH     MS   MSM
## Next.Degree Count 14.000 1.000 1.000 9.000 1.000 6.000 4.000 53.000 1.000
## Ratio              0.014 0.001 0.001 0.009 0.001 0.006 0.004  0.053 0.001
##                      MSW    NDA    NONE    PHD   STM     TC  UBDS  UDDS
## Next.Degree Count 11.000 58.000 378.000 78.000 1.000 22.000 6.000 4.000
## Ratio              0.011  0.058   0.378  0.078 0.001  0.022 0.006 0.004
##                     UMD  UMDS  UNKD
## Next.Degree Count 6.000 2.000 6.000
## Ratio             0.006 0.002 0.006
```

The Next.Degree as a factor variable is too scathered. Many levels only have a single count (ex: MA2, MALS, MSM, BD, etc). We should either group these degrees into few major groups or perhaps create a new factor category for level of study (None, Bachelor, Master, PhD, Other).

2.3.7 AttendenceEvent

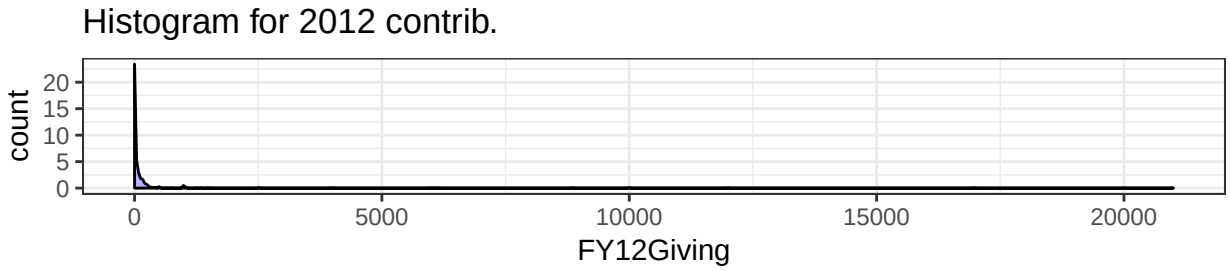
```
row <- xtabs(~AttendenceEvent, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("AttendenceEvent Count",
    "Ratio"))
```

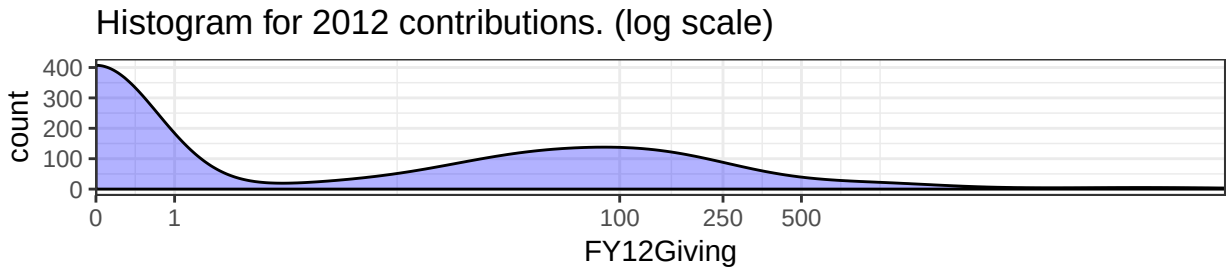
```
##                         Didn.t.Attend Attended
## AttendenceEvent Count         395.000  605.000
## Ratio                           0.395    0.605
```

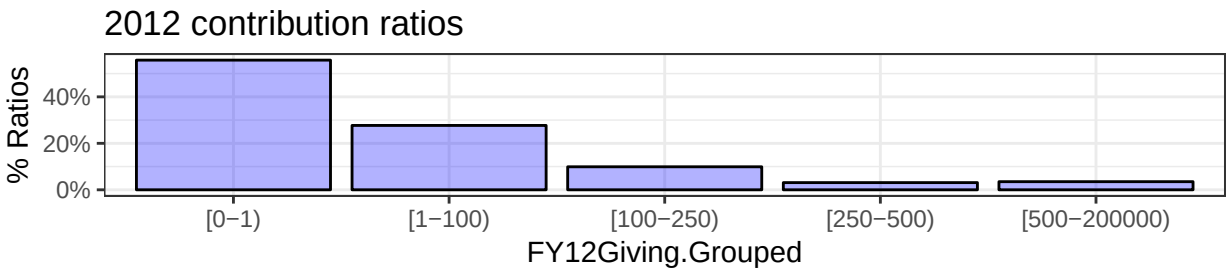
40% of graduates have attended at least one Alumni event organized between 2012 and 2015. This is a very high ratio. Intuitively, we expect a high correlation between this variable and donatios so we'll include this variable in our analysis.

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

```
p1 <- ggplot(givings, aes(x = FY12Giving)) + stat_density(aes(y = ..count..),
    color = "black", fill = "blue", alpha = 0.3) + labs(title = "Histogram for 2012 contrib.") +
    theme_bw()
p2 <- ggplot(givings, aes(x = FY12Giving)) + stat_density(aes(y = ..count..),
    color = "black", fill = "blue", alpha = 0.3) + scale_x_continuous(breaks = c(0,
    1, 100, 250, 500, 2e+05), trans = "log1p", expand = c(0,
    0)) + labs(title = "Histogram for 2012 contributions. (log scale)") +
    scale_y_continuous() + theme_bw()
p3 <- ggplot(givings, aes(x = FY12Giving.Grouped)) + geom_bar(aes(y = (..count..)/sum(..count..)),
    color = "black", fill = "blue", alpha = 0.3) + theme(axis.text.x = element_text(angle = 90,
    hjust = 1)) + scale_y_continuous(labels = percent) + labs(y = "% Ratios",
    title = "2012 contribution ratios") + 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. The third graph captures the ratio of donors in each bracket. It also reduces issues that may arise from high leverage (unique outliers than can skew the model) points.

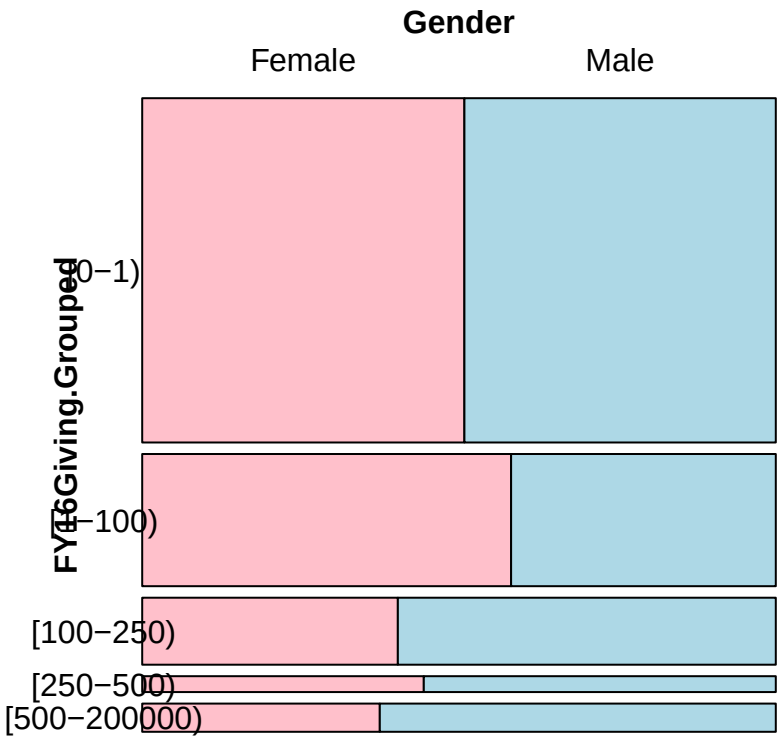
Note (Daghan): I will do some data wrangling to draw all 4 years on the same graphs

2.4 Bivariate Data Analysis

We are going to look at the interaction between 2016 donations level and other variables.

2.4.1 FY16Giving.Grouped vs. Gender

```
mosaic(~FY16Giving.Grouped + Gender, data = givings, highlighting = "Gender",
    highlighting_fill = c("pink", "lightblue"), labeling = labeling_border(rot_labels = c(0,
        0)))
```



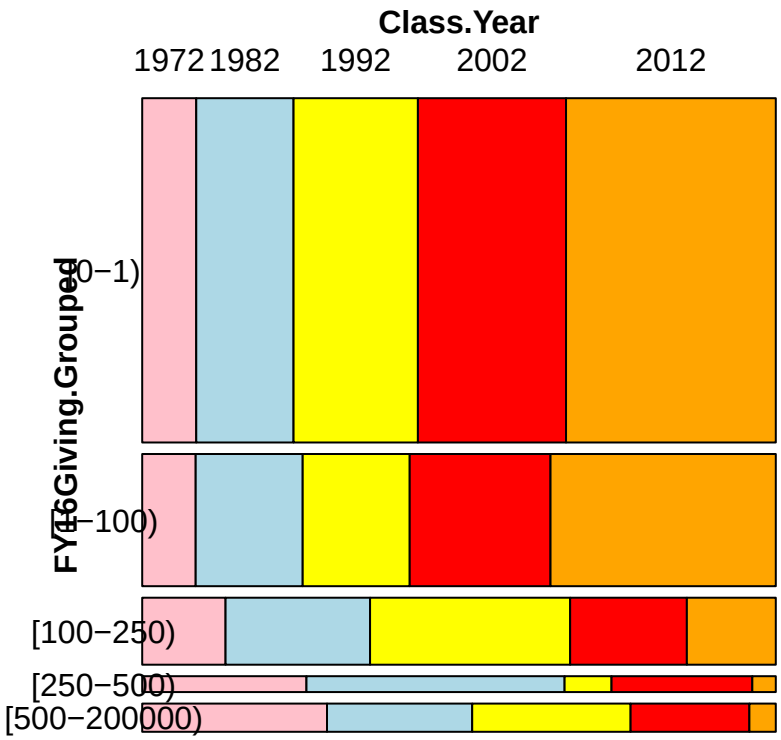

```
xtabs(~FY16Giving.Grouped + Gender, data = givings)
```

```
##                     Gender
## FY16Giving.Grouped Female Male
##        [0-1)          298  288
##        [1-100)        131   94
##        [100-250)       46   68
##        [250-500)       12   15
##        [500-200000)    18   30
```

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. **Gender will be an important factor in our model selection.**

2.4.2 FY16Giving.Grouped vs. Class.Year

```
mosaic(~FY16Giving.Grouped + Class.Year, data = givings, highlighting = "Class.Year",
    highlighting_fill = c("pink", "lightblue", "yellow", "red",
        "orange"), labeling = labeling_border(rot_labels = c(0,
        0)))
```



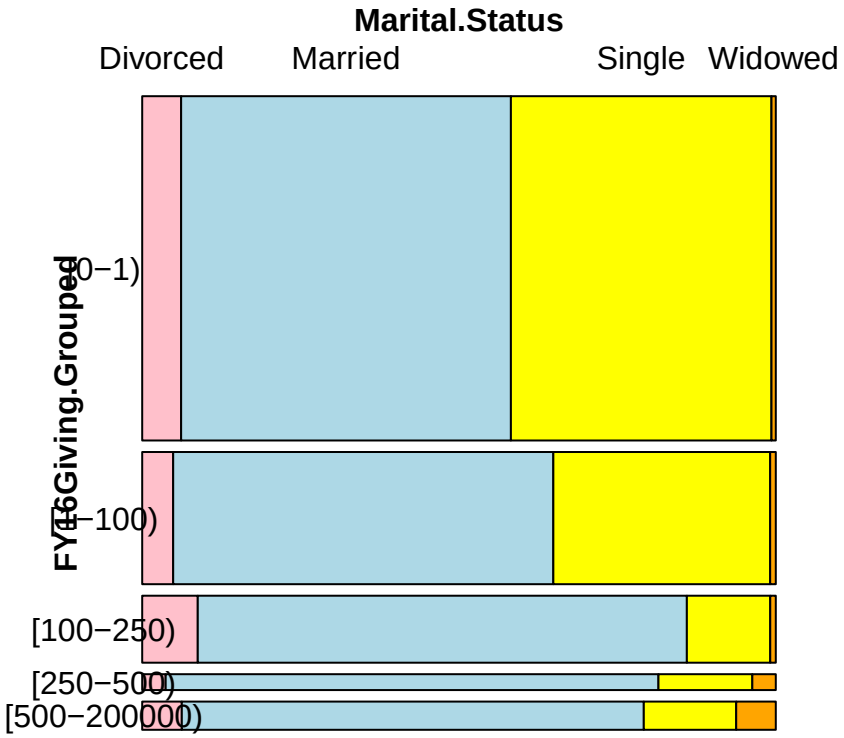

```
xtabs(~FY16Giving.Grouped + Class.Year, data = givings)
```

```
##                    Class.Year
## FY16Giving.Grouped 1972 1982 1992 2002 2012
##        [0-1)         50   90  115  137  194
##        [1-100)       19   38   38   50   80
##        [100-250)     15   26   36   21   16
##        [250-500)      7   11    2    6    1
##        [500-200000)  14   11   12    9    2
```

We observe that older alumni make disportionately bigger donations. What they lack in number (as expected due to age, mortality and other factors), they make up in generousity.

2.4.3 FY16Giving.Grouped vs. Marital.Status

```
mosaic(~FY16Giving.Grouped + Marital.Status, data = givings,
    highlighting = "Marital.Status", highlighting_fill = c("pink",
        "lightblue", "yellow", "orange"), labeling = labeling_border(rot_labels = c(0,
        0)))
```




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

```
##                    Marital.Status
## FY16Giving.Grouped Divorced Married Single Widowed
##        [0-1)               36     305    241        4
##        [1-100)             11     135     77        2
##        [100-250)           10      88     15        1
##        [250-500)            1      21      4        1
##        [500-200000)         3      35      7        3
```

The data is 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.**

2.4.4 FY16Giving.Grouped vs. Major and Grouped.Major

```
# mosaic(~ FY16Giving.Grouped + Major, data = givings,
# highlighting = 'Major', highlighting_fill =c('pink',
# 'lightblue', 'yellow', 'orange'), labeling=
# labeling_border(rot_labels = c(0,0))) xtabs(~
# FY16Giving.Grouped + Major, data = givings) mosaic(~
# FY16Giving.Grouped + Grouped.Major, data = givings,
# highlighting = 'Grouped.Major', highlighting_fill
# =c('pink', 'lightblue', 'yellow', 'orange'), labeling=
# labeling_border(rot_labels = c(0,0))) xtabs(~
# FY16Giving.Grouped + Grouped.Major, data = givings)
```

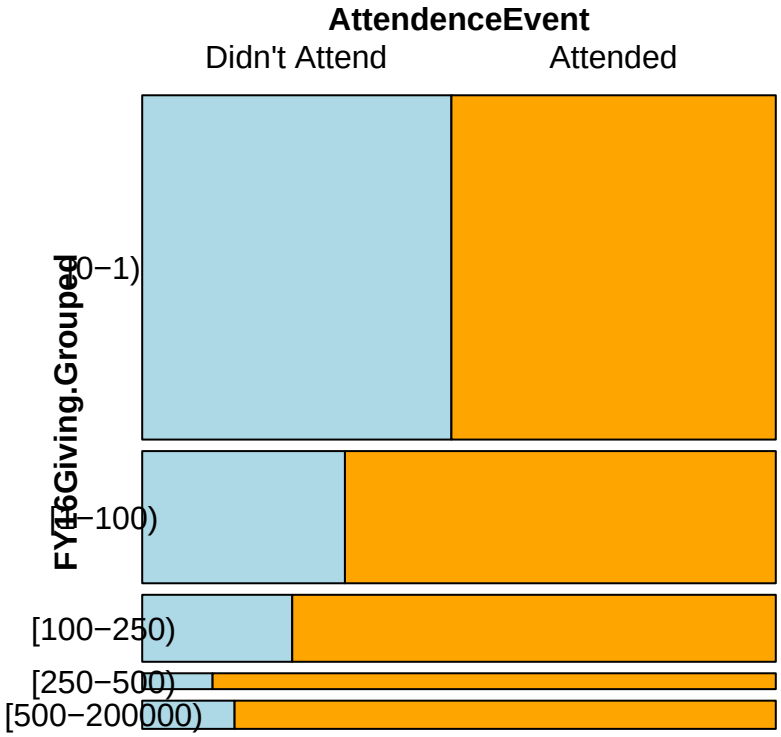
This data will change after Jill's update, skipping for now

2.4.5 FY16Giving.Grouped vs. Next.Degree

This data will change after Jill's update, skipping for now

2.4.6 FY16Giving.Grouped vs. AttendenceEvent

```
mosaic(~FY16Giving.Grouped + AttendenceEvent, data = givings,
    highlighting = "AttendenceEvent", highlighting_fill = c("lightblue",
        "orange"), labeling = labeling_border(rot_labels = c(0,
        0)))
```




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

```
##                    AttendenceEvent
## FY16Giving.Grouped Didn't Attend Attended
##        [0-1)                286      300
##        [1-100)               72      153
##        [100-250)             27       87
##        [250-500)              3       24
##        [500-200000)           7       41
```

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 (85%) have attended an Alumni event. **We'll include AttendenceEvent in our model exploration.**

2.4.7 FY16Giving.Grouped vs. previous years' Donation levels

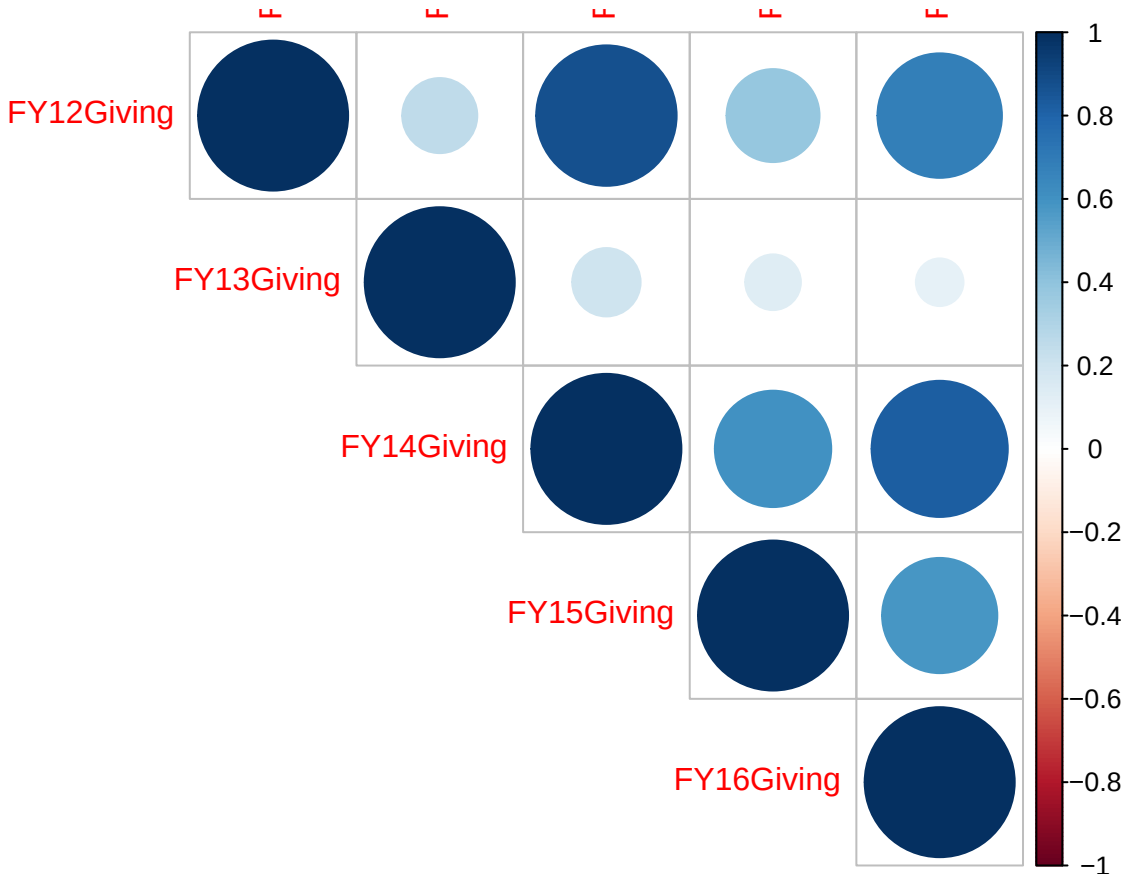
```
library(car)
```

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

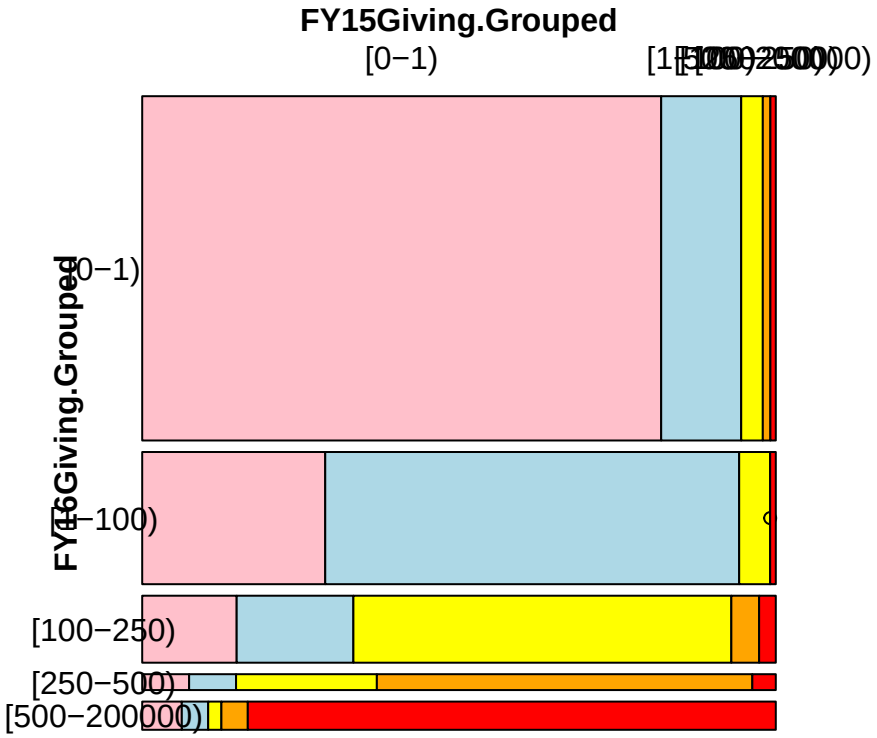
```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## 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")
```




```
mosaic(~FY16Giving.Grouped + FY15Giving.Grouped, data = givings,
    highlighting = "FY15Giving.Grouped", highlighting_fill = c("pink",
        "lightblue", "yellow", "orange", "red"), labeling = labeling_border(rot_labels = c(0,
        0)))
```




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

```
##                     FY15Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##        [0-1)          480      74         20          7             5
##        [1-100)         65     147         11          0             2
##        [100-250)       17      21         68          5             3
##        [250-500)        2       2          6         16             1
##        [500-200000)     3       2          1          2            40
```

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). **So the prior year donation levels are a strong indicator for this year's donation levels.**

Note(Daghan): The above table is for 2016-2015 donations only. I'll add 2014 through 2012 later

2.5 Multivariate Data Analysis and Interactions

Here I am going to explore various combinations and possible interactions. (not done yet)

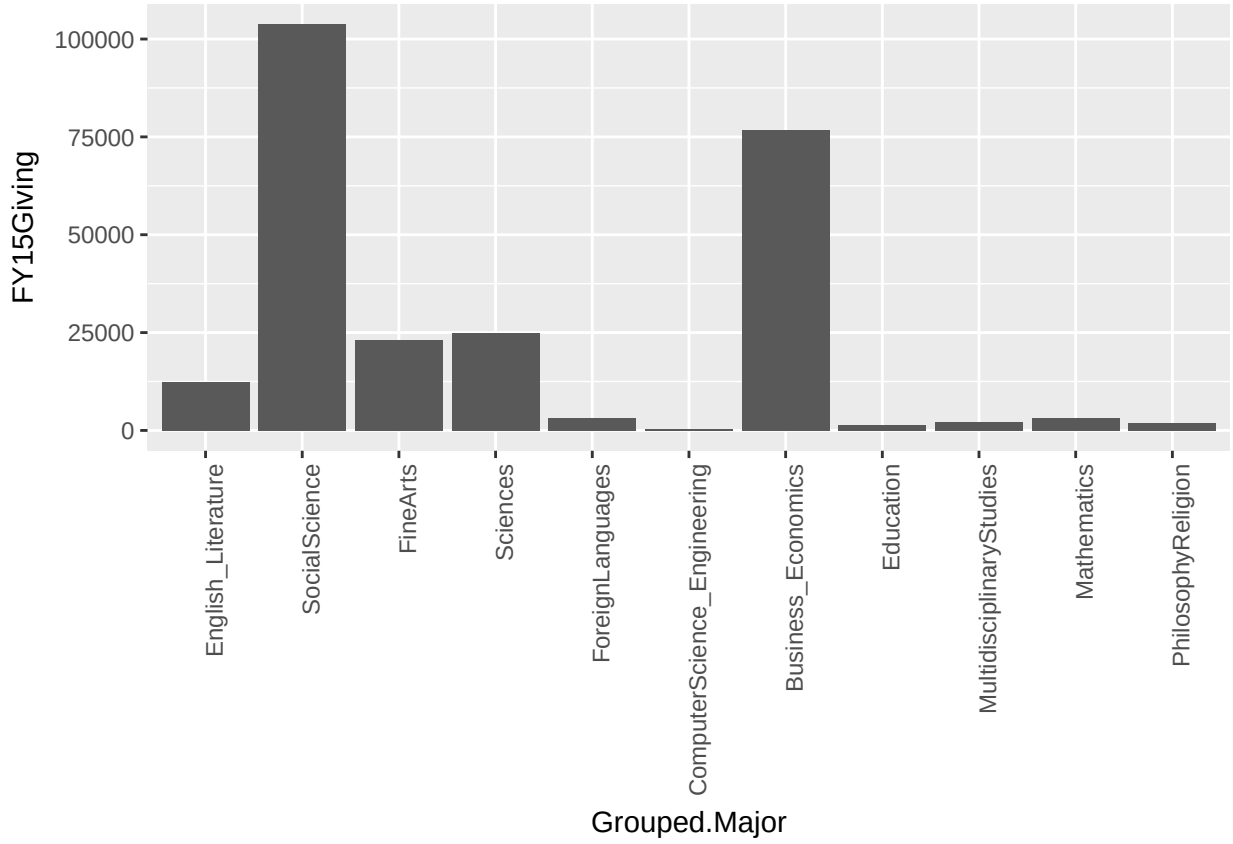
Note (Daghan): Didn't touch the stuff below so you can merge more easily

We need a total donation normalized by number of people (average - median maybe- donation). But considering the above and below charts, it is obvious that business/economics is overrepresented in terms of FY15 giving (consider doing this with averaged 2012, 2014 FY giving when predicting 2015 giving).

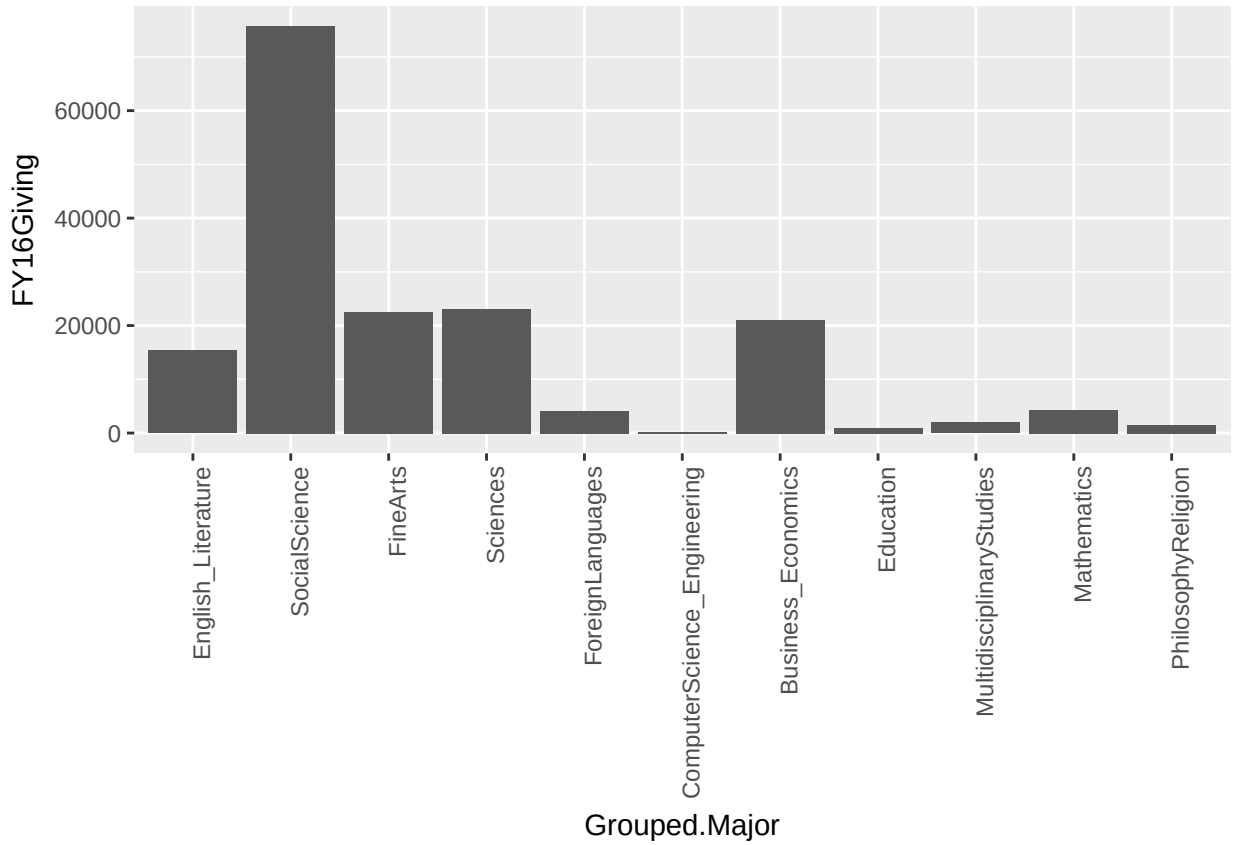
Wow! It seems that the very large donations throw everything off (like which major has higher donations every year fluctuates based on large donations that differ in major from year to year). It seems like FYGiving 2013 is very different from the others because there seems to be a very large donation of >$150,000.

**should i do Median donation by person in each major in each year? need some kind of aggregation - instead maybe do more careful analysis categorical donation variable below.

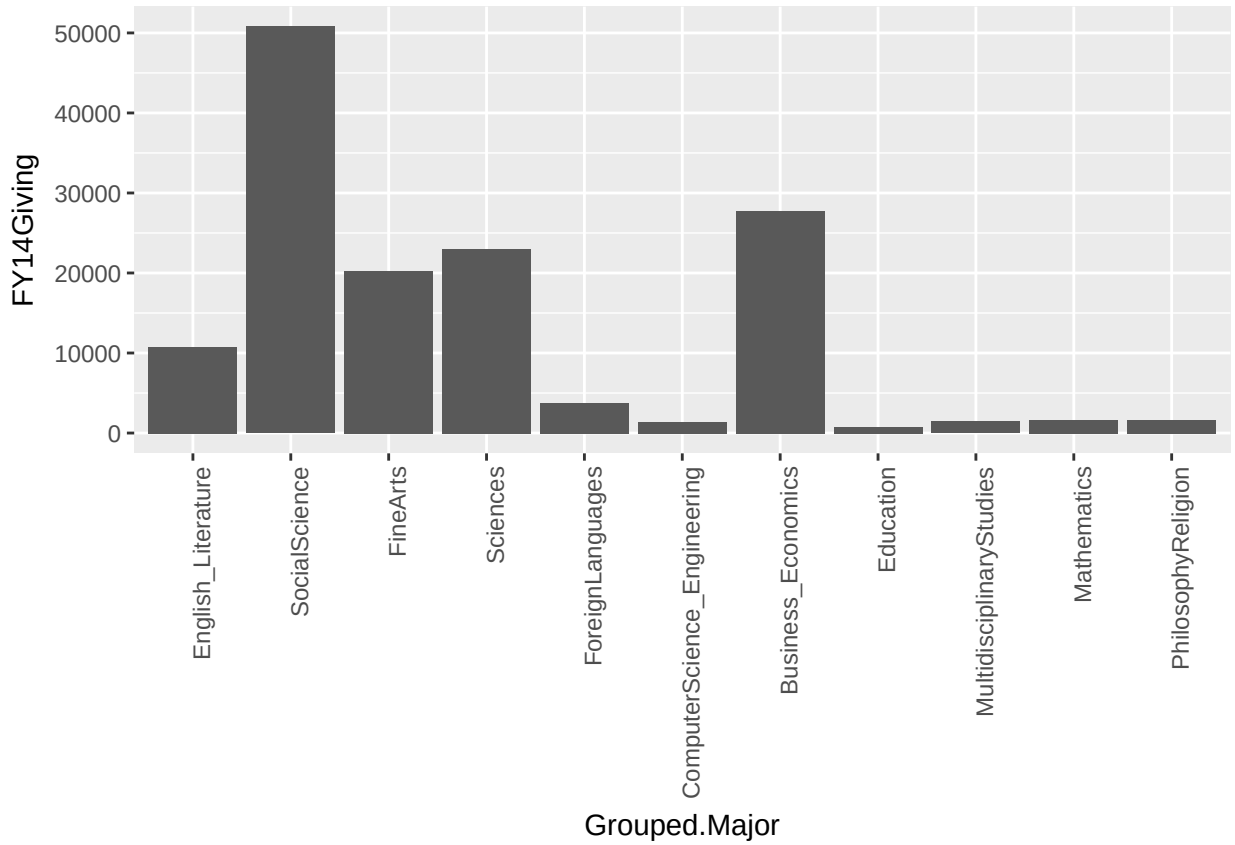
```
ggplot(givings, aes(Grouped.Major, FY15Giving)) + geom_bar(stat = "identity") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



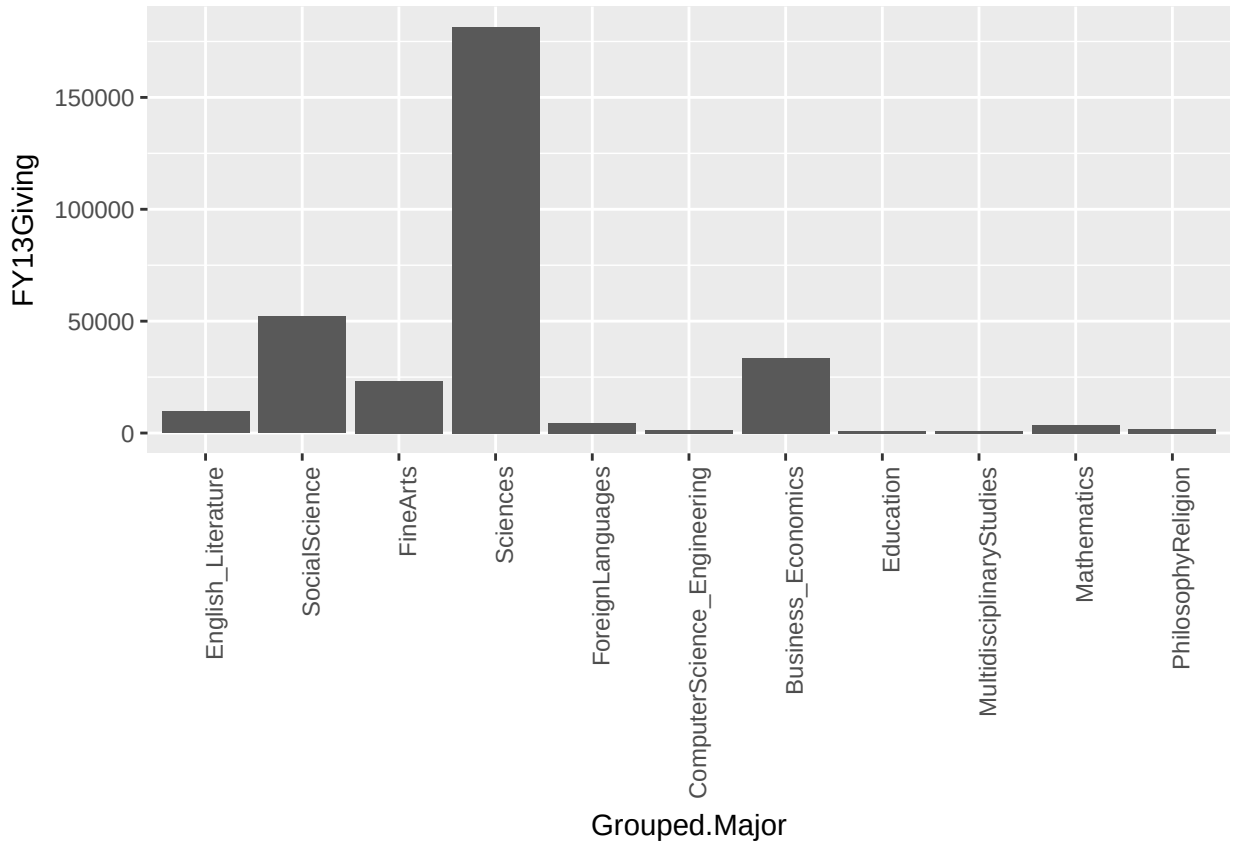

```
ggplot(givings, aes(Grouped.Major, FY16Giving)) + geom_bar(stat = "identity") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



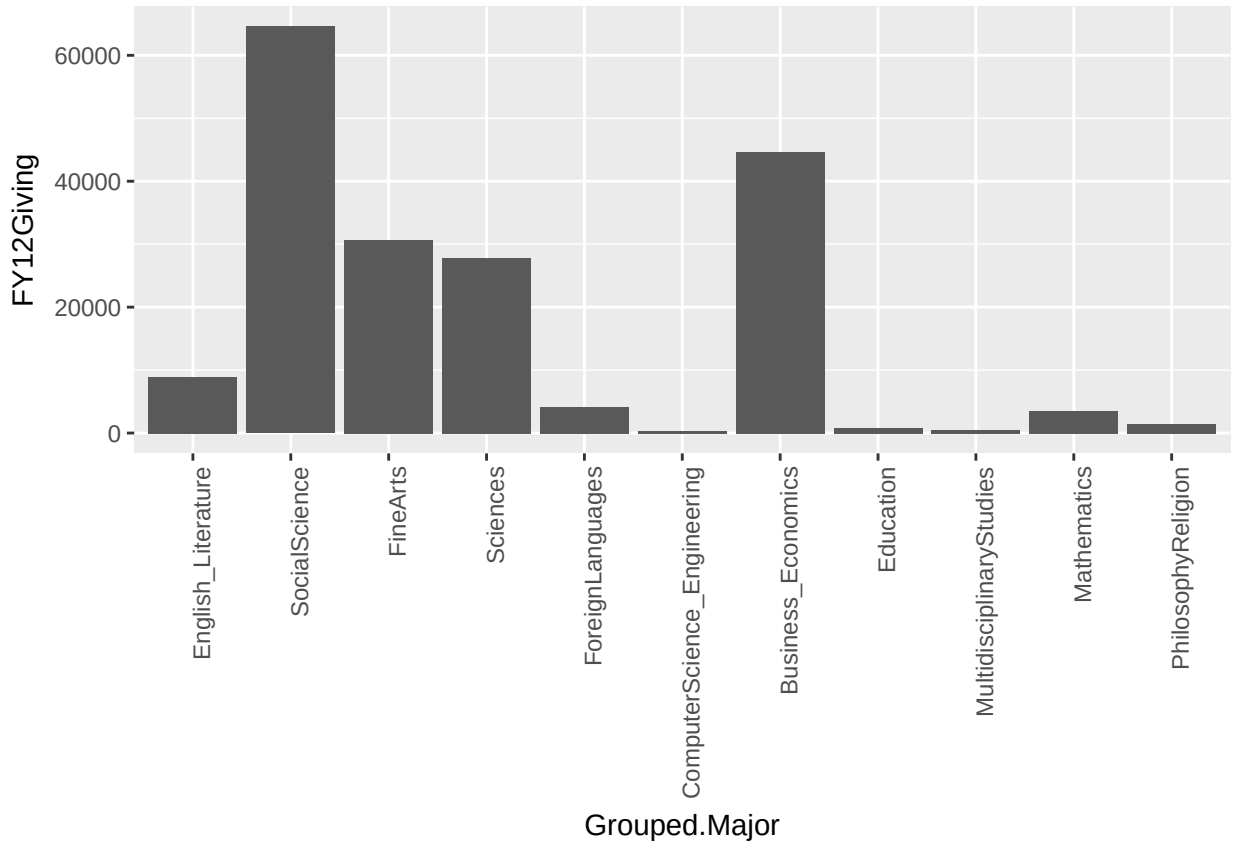

```
ggplot(givings, aes(Grouped.Major, FY14Giving)) + geom_bar(stat = "identity") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
ggplot(givings, aes(Grouped.Major, FY13Giving)) + geom_bar(stat = "identity") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

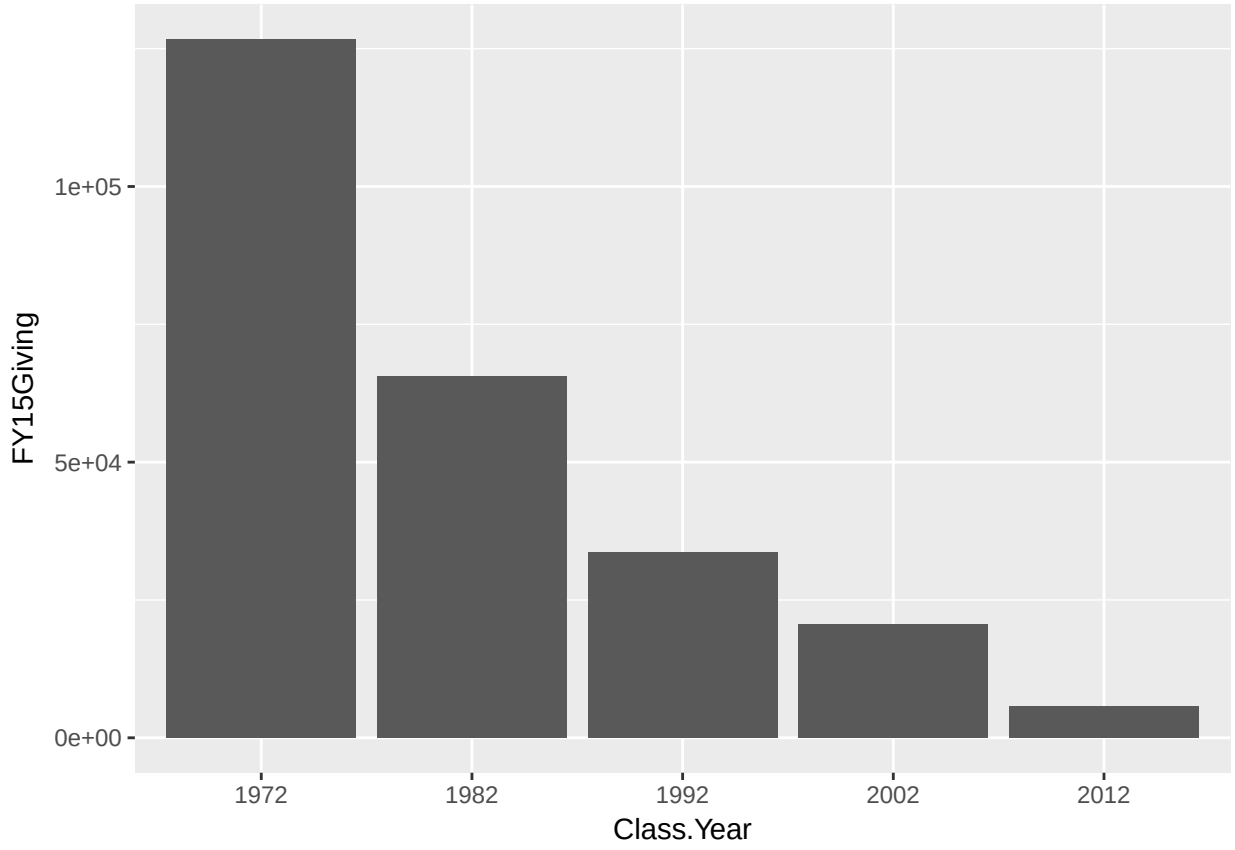



```
ggplot(givings, aes(Grouped.Major, FY12Giving)) + geom_bar(stat = "identity") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Class year affects total donations. People who graduated longer ago are probably older and they seem to donate more total money. It seems like they make more larger donations, but there are fewer people in "older" class years. See contingency table below for grouped giving category observations.

```
ggplot(givings, aes(Class.Year, FY15Giving)) + geom_bar(stat = "identity")
```



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.

```
library(car)
library(corrplot)
M = givings[-c(1:7, 13:17)]
head(M)
```

```
##   FY12Giving FY13Giving FY14Giving FY15Giving FY16Giving
## 1         50         51         51          0          0
## 2          0          0          0          0          0
## 3        100          0        100        100        100
## 4          0          0          0          0          0
## 5          0          0          0          0          0
## 6          0          0          0          0          0
##   FY16Giving.Grouped
## 1              [0-1)
## 2              [0-1)
## 3            [1-100)
## 4              [0-1)
```

```
## 5              [0-1)
## 6              [0-1)
```

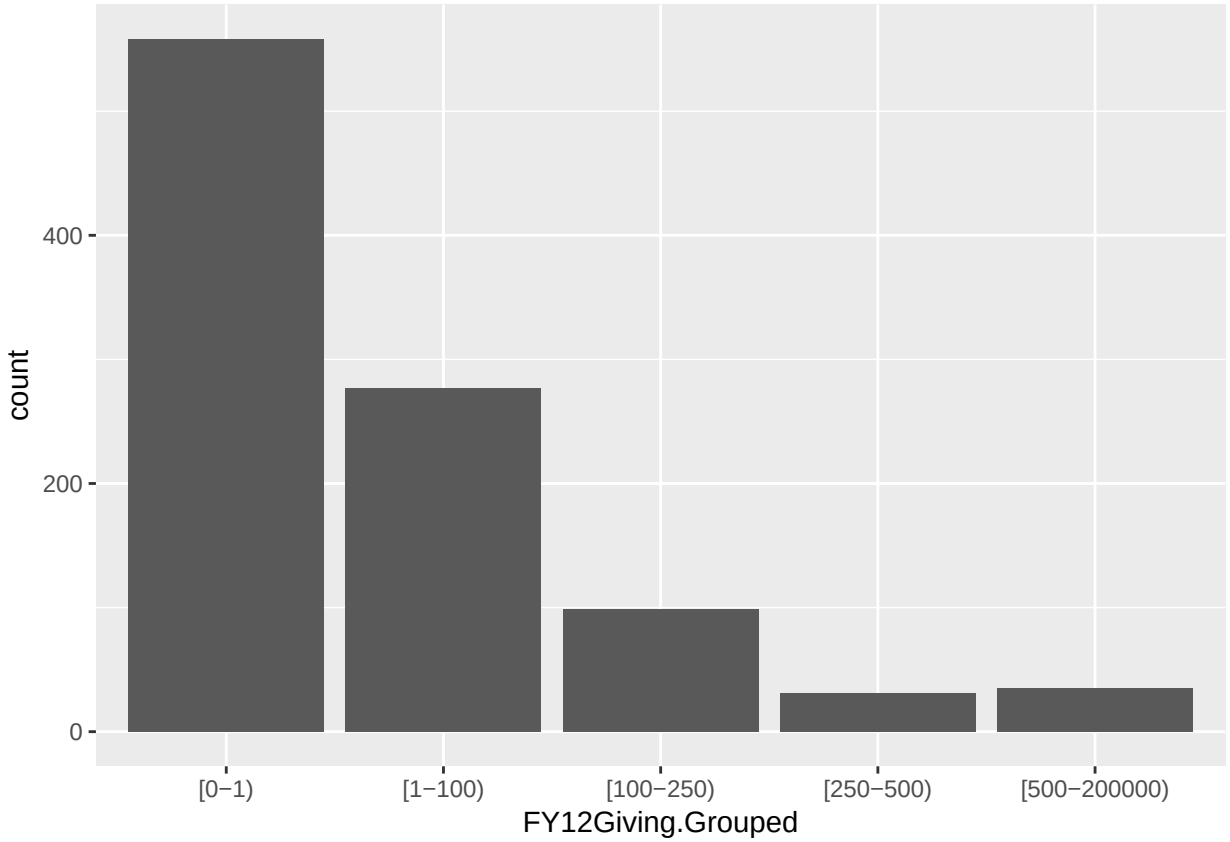
```
# M_corr = cor(M) corrplot(M_corr, method='circle',
# type='upper')
```

Creating grouped (categorical) donation amount variables:

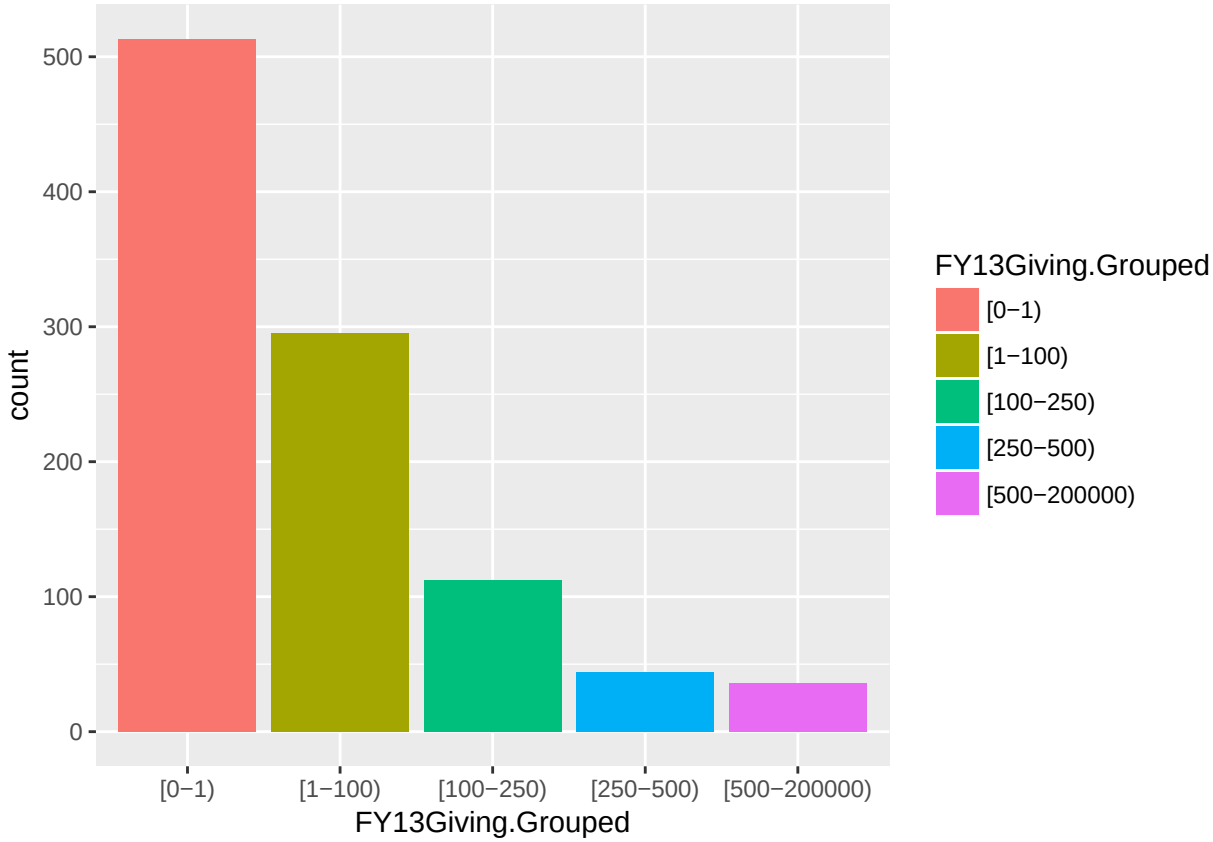
```
givings$FY12Giving.Grouped <- factor(cut(givings$FY12Giving,
    breaks = c(0, 1, 100, 250, 500, 2e+05), labels = c("[0-1)",
        "[1-100)", "[100-250)", "[250-500)", "[500-200000)"),
    include.lowest = TRUE))

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)"),
    include.lowest = TRUE))

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)"),
    include.lowest = TRUE))

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)"),
    include.lowest = TRUE))

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)"),
    include.lowest = TRUE))
```

In general (across years) half of people donate and half do not. Half of people who donate give under a hundred dollars. Around one fith of people who donate (1/10 of all people) give between 100 and 250 dollars. 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.

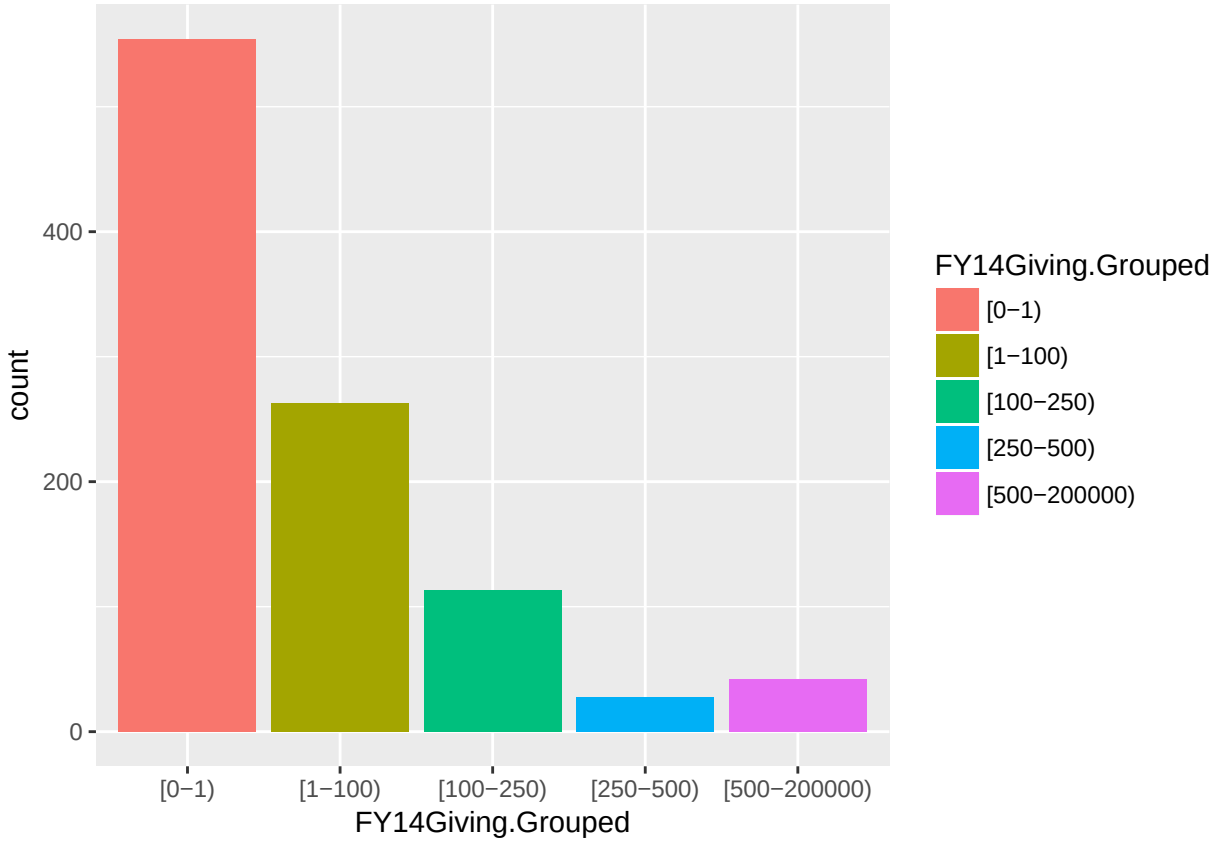
```
ggplot(givings, aes(FY12Giving.Grouped)) + geom_bar()
```



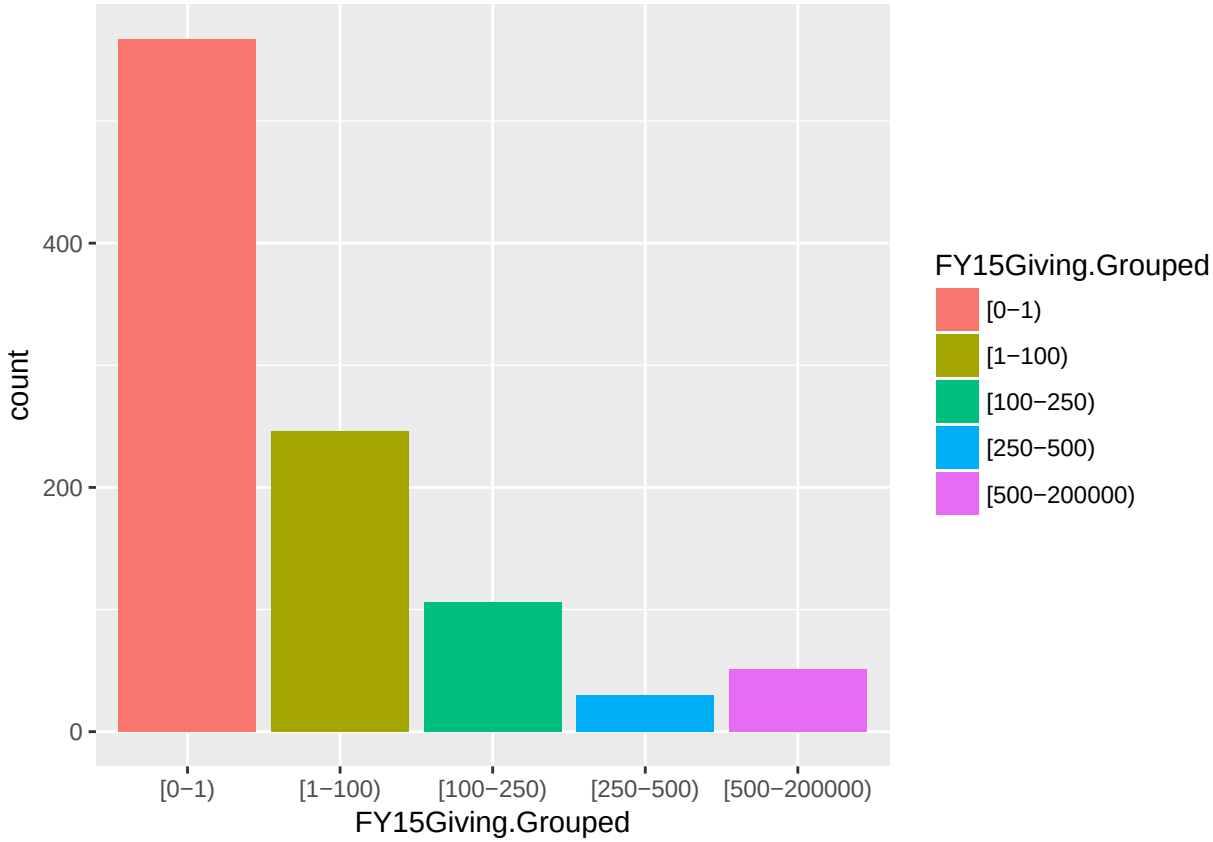

```
ggplot(givings, aes(FY13Giving.Grouped)) + geom_bar(aes(fill = FY13Giving.Grouped))
```



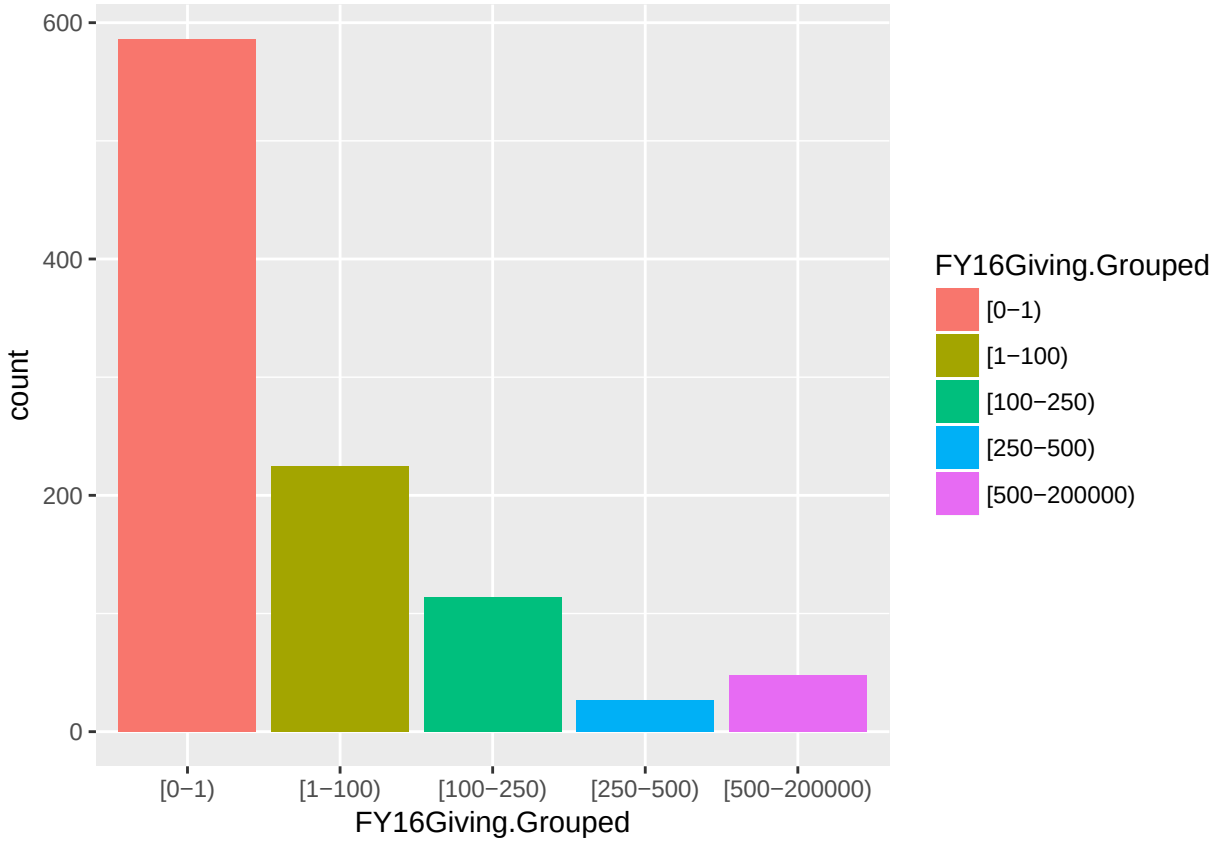

```
ggplot(givings, aes(FY14Giving.Grouped)) + geom_bar(aes(fill = FY14Giving.Grouped))
```




```
ggplot(givings, aes(FY15Giving.Grouped)) + geom_bar(aes(fill = FY15Giving.Grouped))
```




```
ggplot(givings, aes(FY16Giving.Grouped)) + geom_bar(aes(fill = FY16Giving.Grouped))
```

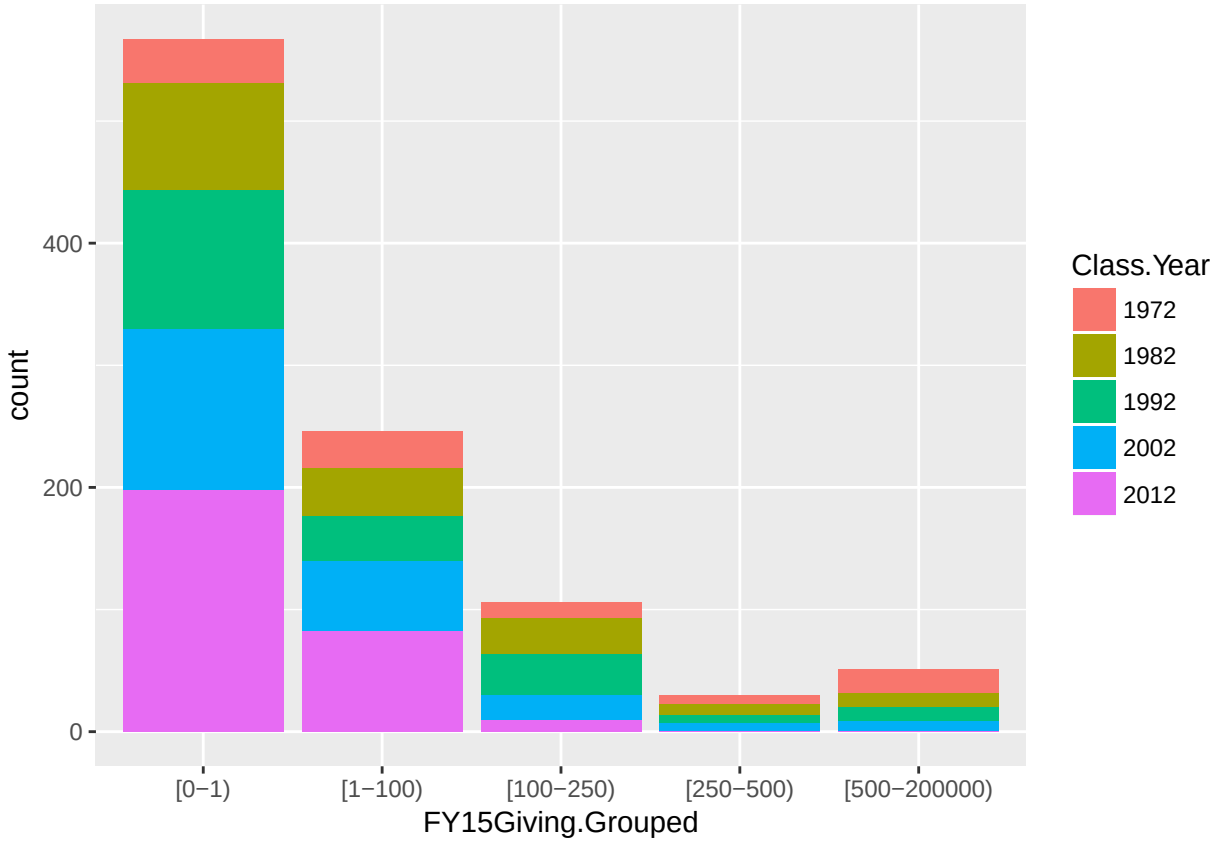


DONATION LEVEL VS. CLASS YEAR

!! Caveat: should make all the bar charts below percent of people in each category that give the particular donation amount - ie. percent of the class of 1972 that donate nothing, percent of the class of 1972 that donate 1-100 dollars. Right now the charts are counts of the categories in the legends and the number of people in each category is different (except for in the case of men and women).

More people from more recent classes do not donate. More people from recent classes donate in lower donation amounts. As graduation year is "older", there are more larger value donations (remember, chart is not normalized to # of people in each category).

```
ggplot(givings, aes(FY15Giving.Grouped)) + geom_bar(aes(fill = Class.Year))
```



As seen in the contingency table below, class year does affect amount donated. The most recent graduates have the highest percent of people who do not donate. Wherease the "oldest"" graduates have the highest amount of people who donate in the higher donation bracket.

The Pearson Chi Square and Likelihood Ratio Tests indicate that we should reject the null hypothesis that Class Year and Donation group are independent of each other (for FY15Giving.Grouped data).

```
Class.year.table <- xtabs(~Class.Year, data = givings)
prop.table(Class.year.table)
```

```
## Class.Year
##  1972  1982  1992  2002  2012
## 0.105 0.176 0.203 0.223 0.293
```

```
# t7 <-xtabs(~FY16Giving.Grouped+Class.Year, data=givings)
# round(t7/rowSums(t7),2)

# t8 <-xtabs(~FY12Giving.Grouped+Class.Year, data=givings)
# round(t8/rowSums(t8),2)

# t9 <-xtabs(~FY13Giving.Grouped+Class.Year, data=givings)
# round(t9/rowSums(t9),2)

# t10 <-xtabs(~FY14Giving.Grouped+Class.Year, data=givings)
# round(t10/rowSums(t10),2)

t11 <- xtabs(~FY15Giving.Grouped + Class.Year, data = givings)
round(t11/rowSums(t11), 2)
```

```
##                    Class.Year
## FY15Giving.Grouped 1972 1982 1992 2002 2012
##        [0-1)        0.06 0.15 0.20 0.23 0.35
##        [1-100)      0.12 0.16 0.15 0.23 0.34
##        [100-250)    0.12 0.27 0.32 0.19 0.09
##        [250-500)    0.23 0.30 0.23 0.20 0.03
##        [500-200000) 0.37 0.24 0.22 0.16 0.02
```

```
assocstats(t11)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 118.84 16        0
## Pearson          115.91 16        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.322
## Cramer's V        : 0.17
```

I created new binary variable to compare number of people who donate versus do not donate for fiscal year 2015.

```
givings$donatedFY15 <- fct_collapse(givings$FY15Giving.Grouped,
    Donated = c("(1-100)", "(100-250)", "(250-500)", "(500-200000)"),
    DidnotDonate = "(0-1)")
```

```
## Warning: Unknown levels in `f`: (1-100), (100-250), (250-500),
## (500-200000), (0-1)
```

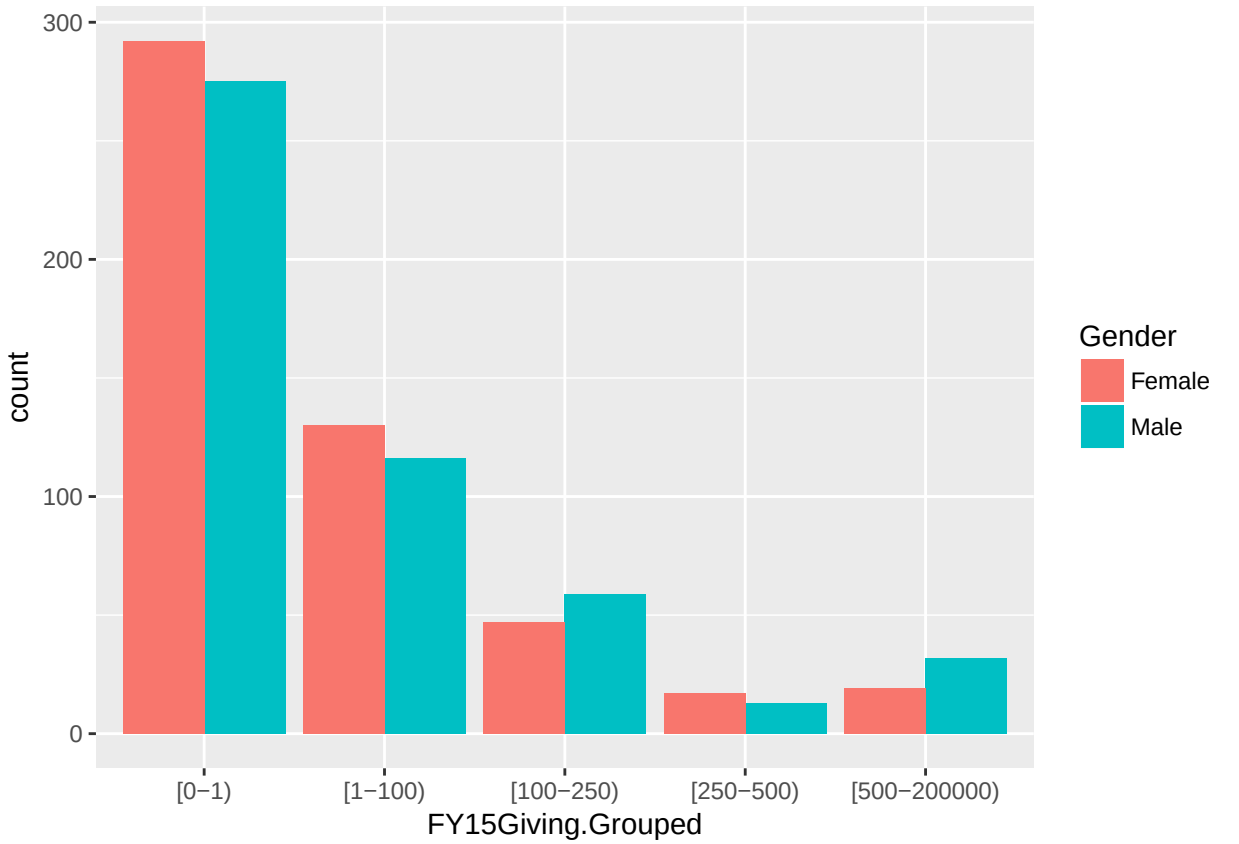
I wanted to conclusively see if the number of people who do and do not donate differ between class years. They do!

```
donated.classyearTable = xtabs(~donatedFY15 + Class.Year, data = givings)
donated.classyearTable
```

```
##               Class.Year
## donatedFY15    1972 1982 1992 2002 2012
##   [0-1)          36   87  114  132  198
##   [1-100)        30   39   37   57   83
##   [100-250)      13   29   34   20   10
##   [250-500)       7    9    7    6    1
##   [500-200000)   19   12   11    8    1
```

DONATION LEVEL VS. GENDER

```
ggplot(givings, aes(FY15Giving.Grouped)) + geom_bar(aes(fill = Gender),
    position = "dodge")
```



FY15 donations are independent of gender according to LR and Pearson Chi square tests. There are clear differences in donation patterns for larger donation categories, but there are few observations in these categories.

```
t1 <- xtabs(~FY15Giving.Grouped + Gender, data = givings)
round(t1/rowSums(t1), 2)
```

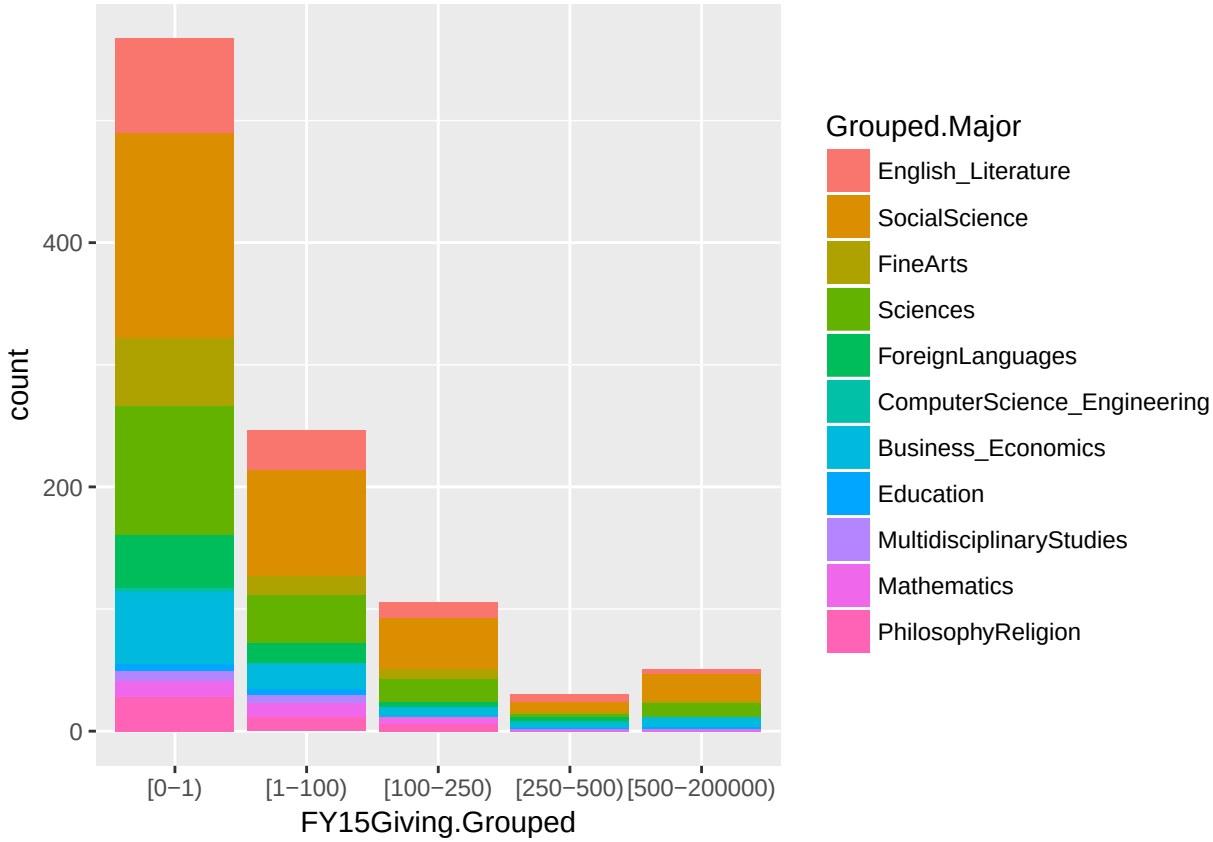
```
##                    Gender
## FY15Giving.Grouped Female Male
##       [0-1)           0.51 0.49
##       [1-100)         0.53 0.47
##       [100-250)       0.44 0.56
##       [250-500)       0.57 0.43
##       [500-200000)    0.37 0.63
```

```
assocstats(t1)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 6.4539  4  0.16772
## Pearson          6.4126  4  0.17038
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.08
## Cramer's V        : 0.08
```

DONATION LEVEL VS. MAJOR

```
ggplot(givings, aes(FY15Giving.Grouped)) + geom_bar(aes(fill = Grouped.Major))
```



Seems marginally significant by LR test that FY15 giving dependent on major. But not significant by Pearson Chi square test. I put it in first pass of models - not convincingly significant!

```
t17 <- xtabs(~FY15Giving.Grouped + Grouped.Major, data = givings)
round(t17/rowSums(t17), 2)
```

```
##                   Grouped.Major
## FY15Giving.Grouped English_Literature SocialScience FineArts Sciences
##       [0-1)                     0.14          0.30     0.10     0.19
##       [1-100)                   0.13          0.35     0.06     0.16
##       [100-250)                 0.12          0.40     0.08     0.18
##       [250-500)                 0.20          0.27     0.07     0.07
##       [500-200000)              0.08          0.43     0.04     0.22
##                   Grouped.Major
## FY15Giving.Grouped ForeignLanguages ComputerScience_Engineering
##       [0-1)                   0.08                        0.01
##       [1-100)                 0.07                        0.00
##       [100-250)               0.04                        0.00
##       [250-500)               0.13                        0.03
##       [500-200000)            0.00                        0.00
##                   Grouped.Major
## FY15Giving.Grouped Business_Economics Education MultidisciplinaryStudies
##       [0-1)                     0.10      0.01                     0.02
```

```
##       [1-100)                   0.08      0.02                   0.03
##       [100-250)                 0.08      0.00                   0.00
##       [250-500)                 0.13      0.03                   0.03
##       [500-200000)              0.18      0.02                   0.02
##                   Grouped.Major
## FY15Giving.Grouped Mathematics PhilosophyReligion
##       [0-1)               0.02               0.05
##       [1-100)             0.04               0.05
##       [100-250)           0.06               0.06
##       [250-500)           0.03               0.00
##       [500-200000)        0.02               0.00
```

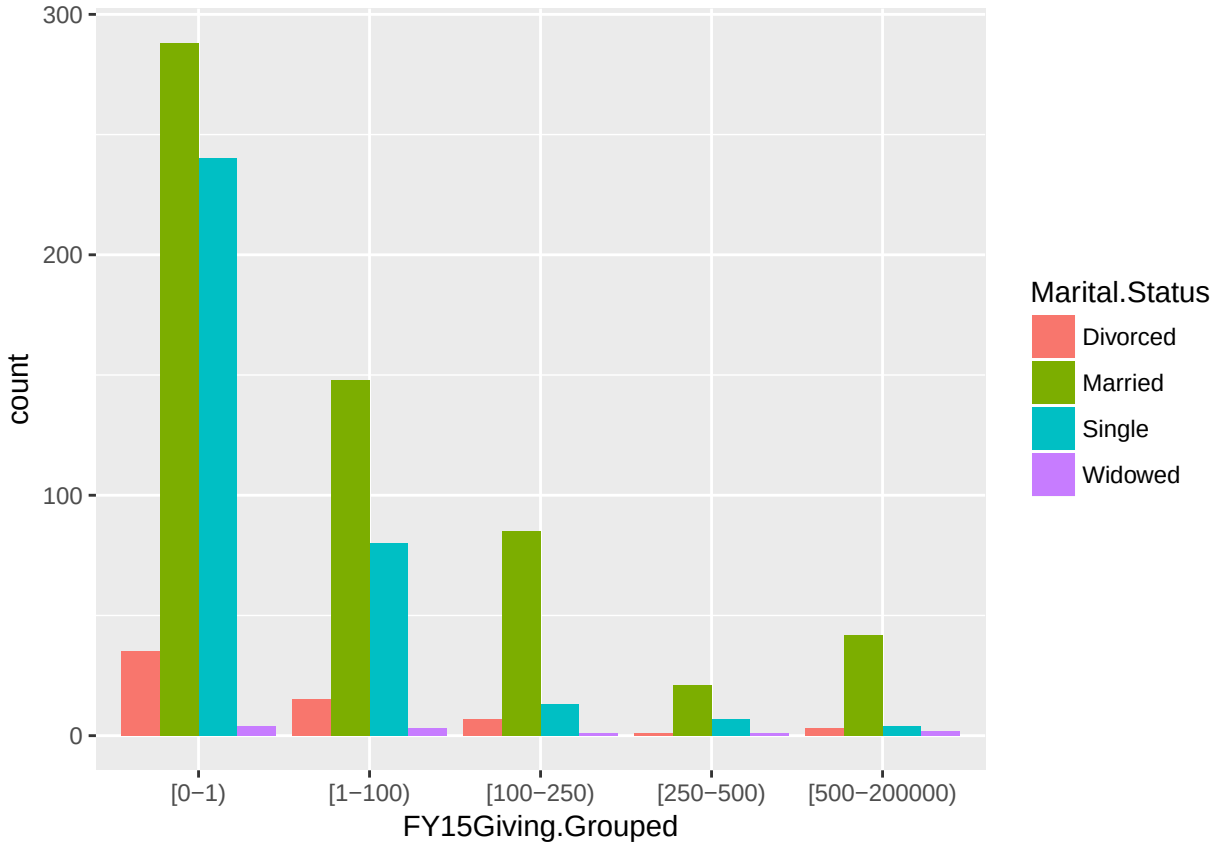
```
assocstats(t17)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 58.422 40 0.030033
## Pearson          51.055 40 0.113015
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.22
## Cramer's V        : 0.113
```

DONATION LEVEL VS. MARITAL STATUS

Marital status seems to contribute to amount donated in all fiscal years. Maried people give more than single people and larger amounts are disproportionately from married people. Divorced and widowed people tend to give less often

```
ggplot(givings) + geom_bar(mapping = aes(FY15Giving.Grouped,
    fill = Marital.Status), position = "dodge")
```



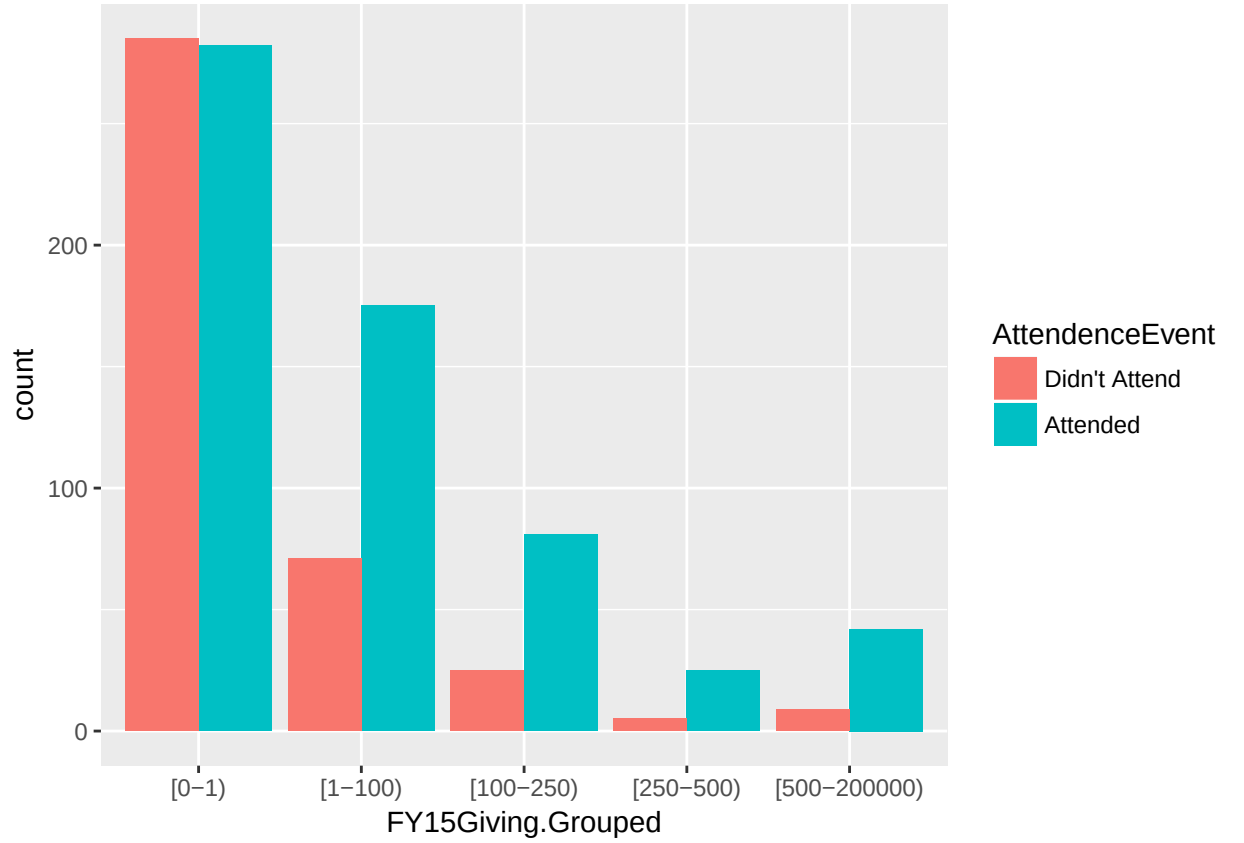
The Pearson Chi Square and LR tests indicate that we should reject the null hypothesis that marital status and donations in each category are independent of each other (for FY15Giving.Grouped data).

```
# t2 <-xtabs(~FY16Giving.Grouped+Marital.Status,
# data=givings) round(t2/rowSums(t2),2)

# t3 <-xtabs(~FY12Giving.Grouped+Marital.Status,
# data=givings) round(t3/rowSums(t3),2)

# t4 <-xtabs(~FY13Giving.Grouped+Marital.Status,
# data=givings) round(t4/rowSums(t4),2)

# t5 <-xtabs(~FY14Giving.Grouped+Marital.Status,
# data=givings) round(t5/rowSums(t5),2)

t6 <- xtabs(~FY15Giving.Grouped + Marital.Status, data = givings)
# round(t6/rowSums(t6),2)
assocstats(t6)
```

```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 69.849 12 3.4178e-10
## Pearson          63.641 12 4.8571e-09
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.245
## Cramer's V        : 0.146
```

DONATION LEVEL VS. ATTENDANCE EVENTS

```
ggplot(givings) + geom_bar(mapping = aes(FY15Giving.Grouped,
    fill = AttendenceEvent), position = "dodge")
```



People who attended events are more likely to give in every donation level category (FY15 dependent on attendance of events).

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

# t13 <-xtabs(~FY12Giving.Grouped+AttendenceEvent,
# data=givings) round(t13/rowSums(t13),2)

# t14 <-xtabs(~FY13Giving.Grouped+AttendenceEvent,
# data=givings) round(t14/rowSums(t14),2)

# t15 <-xtabs(~FY14Giving.Grouped+AttendenceEvent,
# data=givings) round(t15/rowSums(t15),2)

t16 <- xtabs(~FY15Giving.Grouped + AttendenceEvent, data = givings)
round(t16/rowSums(t16), 2)
```

```
##                   AttendenceEvent
## FY15Giving.Grouped Didn't Attend Attended
##       [0-1)                0.50     0.50
##       [1-100)              0.29     0.71
```

```
##       [100-250)             0.24     0.76
##       [250-500)             0.17     0.83
##       [500-200000)          0.18     0.82
```

```
assocstats(t16)
```

```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 69.834  4 2.4647e-14
## Pearson          67.114  4 9.2260e-14
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.251
## Cramer's V        : 0.259
```

Wanted to check in general if attending events increased chance of donating (indepent of size of donation). It does.

```
donated.attendanceTable = xtabs(~donatedFY15 + AttendenceEvent,
    data = givings)
donated.attendanceTable
```

```
##               AttendenceEvent
## donatedFY15    Didn't Attend Attended
##   [0-1)                  285      282
##   [1-100)                 71      175
##   [100-250)               25       81
##   [250-500)                5       25
##   [500-200000)             9       42
```

```
prop.table(donated.attendanceTable)
```

```
##               AttendenceEvent
## donatedFY15    Didn't Attend Attended
##   [0-1)                0.285    0.282
##   [1-100)              0.071    0.175
##   [100-250)            0.025    0.081
##   [250-500)            0.005    0.025
##   [500-200000)         0.009    0.042
```

DONATION LEVEL VS. DONATION LEVEL IN PREVIOUS YEAR

Fiscal year 2015 donation levels are very obviously not independent of 2014 giving levels.

```
t0 <- xtabs(~FY15Giving.Grouped + FY14Giving.Grouped, data = givings)
round(t0/rowSums(t0), 2)
```

```
##                   FY14Giving.Grouped
## FY15Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)         0.80    0.15      0.04      0.01         0.00
##       [1-100)       0.28    0.67      0.04      0.01         0.00
##       [100-250)     0.18    0.11      0.67      0.02         0.02
##       [250-500)     0.17    0.03      0.20      0.57         0.03
##       [500-200000)  0.10    0.02      0.08      0.06         0.75
```

```
assocstats(t0)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio  839.59 16        0
```

```
## Pearson          1647.95 16        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.789
## Cramer's V        : 0.642
```

Preliminary Modeling

```
# 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(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
model1 <- polr(formula = FY15Giving.Grouped ~ Class.Year + Grouped.Major +
    Marital.Status + AttendenceEvent + FY14Giving.Grouped, data = givings,
    method = "logistic", Hess = TRUE)

summary(model1)
```

```
## Call:
## polr(formula = FY15Giving.Grouped ~ Class.Year + Grouped.Major +
##     Marital.Status + AttendenceEvent + FY14Giving.Grouped, data = givings,
##     Hess = TRUE, method = "logistic")
##
## Coefficients:
##                                              Value Std. Error   t value
## Class.Year1982                           -0.7093961     0.2707 -2.620370
## Class.Year1992                           -0.9156060     0.2717 -3.369593
## Class.Year2002                           -0.8331680     0.2736 -3.045228
## Class.Year2012                           -0.9270929     0.2774 -3.342253
## Grouped.MajorSocialScience                0.3655454     0.2274  1.607315
## Grouped.MajorFineArts                    -0.3304133     0.3300 -1.001388
## Grouped.MajorSciences                    -0.1389501     0.2611 -0.532240
## Grouped.MajorForeignLanguages            -0.2818735     0.3438 -0.819760
## Grouped.MajorComputerScience_Engineering  0.3430542     1.2490  0.274672
## Grouped.MajorBusiness_Economics          -0.0006069     0.3047 -0.001992
```

```
## Grouped.MajorEducation                    0.1460583     0.6529  0.223722
## Grouped.MajorMultidisciplinaryStudies     0.6734920     0.5213  1.292001
## Grouped.MajorMathematics                  0.6725754     0.3932  1.710664
## Grouped.MajorPhilosophyReligion           0.2092423     0.3711  0.563894
## Marital.StatusMarried                     0.2534578     0.3120  0.812313
## Marital.StatusSingle                     -0.0472717     0.3407 -0.138756
## Marital.StatusWidowed                     0.2101301     0.7123  0.294989
## AttendenceEventAttended                   0.5235779     0.1581  3.312523
## FY14Giving.Grouped[1-100)                 1.8172014     0.1651 11.007321
## FY14Giving.Grouped[100-250)               3.7885789     0.2560 14.797157
## FY14Giving.Grouped[250-500)               5.4430930     0.4474 12.165445
## FY14Giving.Grouped[500-200000)            8.8663695     0.6388 13.878800
##
## Intercepts:
##                         Value   Std. Error t value
## [0-1)|[1-100)            1.2250  0.4086     2.9983
## [1-100)|[100-250)        3.3666  0.4266     7.8925
## [100-250)|[250-500)      5.4893  0.4680    11.7303
## [250-500)|[500-200000)   6.8422  0.5236    13.0675
##
## Residual Deviance: 1596.823
## AIC: 1648.823
```

```
Anova(model1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY15Giving.Grouped
##                     LR Chisq Df Pr(>Chisq)
## Class.Year             13.42  4  0.0093964 **
## Grouped.Major          15.83 10  0.1046168
## Marital.Status          3.45  3  0.3268428
## AttendenceEvent        11.12  1  0.0008534 ***
## FY14Giving.Grouped    544.49  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model1b <- polr(formula = FY15Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + FY14Giving.Grouped, data = givings, method = "logistic",
    Hess = TRUE)

summary(model1b)
```

```
## Call:
## polr(formula = FY15Giving.Grouped ~ Class.Year + Marital.Status +
##     AttendenceEvent + FY14Giving.Grouped, data = givings, Hess = TRUE,
##     method = "logistic")
##
## Coefficients:
##                                  Value Std. Error t value
## Class.Year1982                -0.69820     0.2607 -2.6779
## Class.Year1992                -0.95841     0.2594 -3.6944
## Class.Year2002                -0.89040     0.2597 -3.4289
## Class.Year2012                -0.99215     0.2652 -3.7414
## Marital.StatusMarried          0.35589     0.3123  1.1394
```

```
## Marital.StatusSingle            0.08707     0.3407  0.2556
## Marital.StatusWidowed           0.16109     0.6769  0.2380
## AttendenceEventAttended         0.57275     0.1567  3.6554
## FY14Giving.Grouped[1-100)       1.77963     0.1629 10.9269
## FY14Giving.Grouped[100-250)     3.71902     0.2533 14.6806
## FY14Giving.Grouped[250-500)     5.38642     0.4411 12.2106
## FY14Giving.Grouped[500-200000)  8.76459     0.6352 13.7990
##
## Intercepts:
##                         Value   Std. Error t value
## [0-1)|[1-100)            1.1989  0.3571     3.3574
## [1-100)|[100-250)        3.3073  0.3762     8.7905
## [100-250)|[250-500)      5.4125  0.4229    12.7993
## [250-500)|[500-200000)   6.7597  0.4832    13.9898
##
## Residual Deviance: 1612.653
## AIC: 1644.653
```

```
Anova(model1b)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY15Giving.Grouped
##                     LR Chisq Df Pr(>Chisq)
## Class.Year             16.87  4  0.0020444 **
## Marital.Status          3.51  3  0.3195330
## AttendenceEvent        13.58  1  0.0002282 ***
## FY14Giving.Grouped    540.21  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

```
## # weights:  120 (92 variable)
## initial  value 1609.437912
## iter  10 value 843.008425
## iter  20 value 693.932143
## iter  30 value 680.087189
## iter  40 value 679.049338
## iter  50 value 678.820096
## iter  60 value 678.734454
## iter  70 value 678.695055
## iter  80 value 678.687982
## iter  90 value 678.687439
## final  value 678.687321
## converged
```

```
Anova(model2)
```

```
## Analysis of Deviance Table (Type II tests)
```

```
##
## Response: FY15Giving.Grouped
##                     LR Chisq Df Pr(>Chisq)
## Class.Year             36.77 16   0.002262 **
## Grouped.Major          53.60 40   0.073699 .
## Marital.Status         12.11 12   0.436814
## AttendenceEvent        14.57  4   0.005685 **
## FY14Giving.Grouped    683.03 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
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 770.401719
## iter  20 value 711.289384
## iter  30 value 706.027509
## iter  40 value 705.510178
## iter  50 value 705.492502
## iter  60 value 705.488753
## final  value 705.488020
## converged
```

```
Anova(model2a)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY15Giving.Grouped
##                     LR Chisq Df Pr(>Chisq)
## Class.Year             39.83 16  0.0008255 ***
## Marital.Status         12.51 12  0.4059224
## AttendenceEvent        17.55  4  0.0015091 **
## FY14Giving.Grouped    687.12 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

```
# here we predict estimated probability of being in a certain
# amount donated category based on values of explanatory
# variables we choose - can we feed in the FY15 dataframe to
# get out predictions for 2016?. Or should we just give
# specific combinations of variables that we think are the
# most intersting/enlightening?

# We probably compare and contrast predictions for
# multinomial regression vs ordinal regression like he did in
# the LiveSession04_v3_JY document.
```

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?