Lab 02

Alex Yang, John Kenney, Ram Balasubramanian

Oct 12, 2017

SECTION - 1 INTRODUCTION & KEY RESULTS

Problem Introduction:

We have been hired by a Private University to identify who among their Alumni are most likely to contribute towards the University's foundation in future years. The university has provided us with data on past contributions from graduates - data includes some demogracramerc information (like gender, marital status), university specific information (graduation year, major of studies), and some information on how "connected"" an Alumnus is to the school (Alumni event attendance, historical contributions).

1.1 HIGH LEVEL DESCRIPTION OF MODELING APPROACH:

We have taken two approaches to the problem (named Beta-Hat and Y-Hat):
Approach "Beta-Hat": We will treat the problem as a "explanation" problem ($\hat{\beta}$). The goal here is to figure out if and how much certain aspects of a person and their association with the university determines how much they will contribute to the university's foundation. We will develop a regression model that considers the 2016 contributions as a variable that depends on one or more of the other data elements that have been provided. The regression coefficients can then be interpreted as a measure of how much each aspect of a person influences their contributions.

Approach "Y-Hat":
We will treat the problem as a "prediction" problem ($\hat{y}$ problem). Given all the data we have about a person and their past contributions, can we predict how much they will contribute in the future. We will develop a model that aims to predict the 2016 contribution amounts for each person. To evaluate the efficacy of our models, we will split the data into a "training" set and a "test" set. We will use the training data to estimate parameters for our prediction model and evaluate our model's prediction accuracy using the test set.

1.2 KEY RESULTS AND TECHNIQUES USED:

We will complete this section once we are done with the modeling work.

SECTION 2 - DATA EXAMINATION AND EDA:

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
x<-c("car","dplyr","Hmisc","ggplot2","gridExtra","effsize","aod","mcprofile","MASS","vcd","data.ta
invisible(lapply(x, require, character.only = TRUE))
```

```
dt <- fread("lab2data.csv")
# describe(dt) #This takes 2.5 pages on its own
```

2.1 Brief Description of Data Available:

We have data for 1000 past graduates of the University. There are 12 variables provided for each Alumnus. They are: 1. V1: Identifier for each record (Alumnus)
2. Gender: M/F, roughly 50/50 in the sample provided. 3. Class.Year: Appears like the "decade" of the graduating year. Goes from 1972 - 2012. We will assume 1972 represents students graduating from 1963-1972; 1982 represents students graduating from 1973 to 1982 etc.
4. Marital.Status: Has 4 categories - coded D,M,S,W. We will assume it means Divorced, Married, Single, Widowed with over 90% in the "married" or "single" categories.
5. Major: There are 45 majors represented in the sample. History, English, Biology & Economics are the top 4 representing about 37% of the sample.
6. Next.Degree: We assume this means what the alumnus went on to do after graduating from the university. 38% shows "None" implying they did not pursue another degree. The remainder (62%) seems rather high for this metric.
7. AttendenceEvent: Indicates whether the alumnus attended an alumni event between 2012 and 2015. If we choose to use this variable to model "Giving" we should probably not use it to model 2012-2015 Giving
8. FYGiving: There are 5 of these variables named FY12 - FY16 representing full year 2012 through full year 2016 contribution from the alumnus. There are some "outliers" (e.g. $161,500 in 2013) in the data that we may need to be on the lookout for.

We do not have any missing values in the data; and there do not seem to be an obvious "data cleaning" that needs to be conducted. We will conduct an Exploratory Data Analysis next.

What are the important variables we want to include in our discussion here? What would we suppose would be meaningful? What can we omit?

```
# View the contents of Major and Next Degree - to identify if
# there are any
majortable = as.data.frame(round(prop.table(table(dt$Major)),
    2))
degtable = as.data.frame(round(prop.table(table(dt$Next.Degree)),
    2))
```

1.3 Create new variables:

Let's group the yearly contributions by the categories that the university is interested in; Classify the "next degree" variable into 0 (representing "none") and 1 (representing there was some next-degree). Create indicator variables for each year for giver(1) or not a giver(0). For each alumnus let's also count the number of years they have given between 2012 and 2015.

```
dt$FY16GivingCat <- cut(dt$FY16Giving, c(0, 1, 100, 250, 500,
    2e+05), right = FALSE)
dt$FY15 <- cut(dt$FY15Giving, c(0, 1, 100, 250, 500, 2e+05),
    right = FALSE)
dt$FY14 <- cut(dt$FY14Giving, c(0, 1, 100, 250, 500, 2e+05),
    right = FALSE)
dt$FY13 <- cut(dt$FY13Giving, c(0, 1, 100, 250, 500, 2e+05),
```

```
    right = FALSE)
dt$FY12 <- cut(dt$FY12Giving, c(0, 1, 100, 250, 500, 2e+05),
    right = FALSE)

# create an indicator for 'giver' and 'non giver' for each
# year.
dt$Giver16 = as.integer(dt$FY16Giving > 0)
dt$Giver15 = as.integer(dt$FY15Giving > 0)
dt$Giver14 = as.integer(dt$FY14Giving > 0)
dt$Giver13 = as.integer(dt$FY13Giving > 0)
dt$Giver12 = as.integer(dt$FY12Giving > 0)
dt$YearsGiven = dt$Giver12 + dt$Giver13 + dt$Giver15 + dt$Giver15

# Create identifier for next degree (1) or none (0)
dt$NextDeg = 1 - as.integer((dt$Next.Degree == "NONE"))

# Group majors by broad categories

dt$MajorCat = ifelse(dt$Major %in% c("American Studies", "Art",
    "Chinese", "Classics", "Comparative Literature", "English",
    "English-Journalism", "French", "German", "History", "Independent",
    "Music", "Philosophy", "Philosophy-Religion", "Physical Education",
    "Religious Studies", "Russian", "Spanish", "Speech (Drama, etc.)",
    "Theatre"), "HUM_ART", ifelse(dt$Major %in% c("Biology",
    "Chemistry", "Computer Science", "Engineering", "General Science",
    "General Science-Biology", "General Science-Chemistry", "General Science-Math",
    "General Science-Physics", "Mathematics", "Mathematics-Physics",
    "Physics", "Zoology"), "STEM", ifelse(dt$Major %in% c("Economics",
    "Economics-Regional Stds.", "Sociology", "Pyschology", "Pol. Sci.-Regional Stds.",
    "Sociology-Anthropology", "Political Science", "Anthropology",
    "Economics-Business"), "SOCIAL_SCIENCE", "OTHER")))

dt$MS = factor(dt$Marital.Status)
dt$Class = factor(dt$Class.Year)
names(dt)[names(dt) == "AttendenceEvent"] = "Event"
```

2.2 Exploratory Data Analysis:

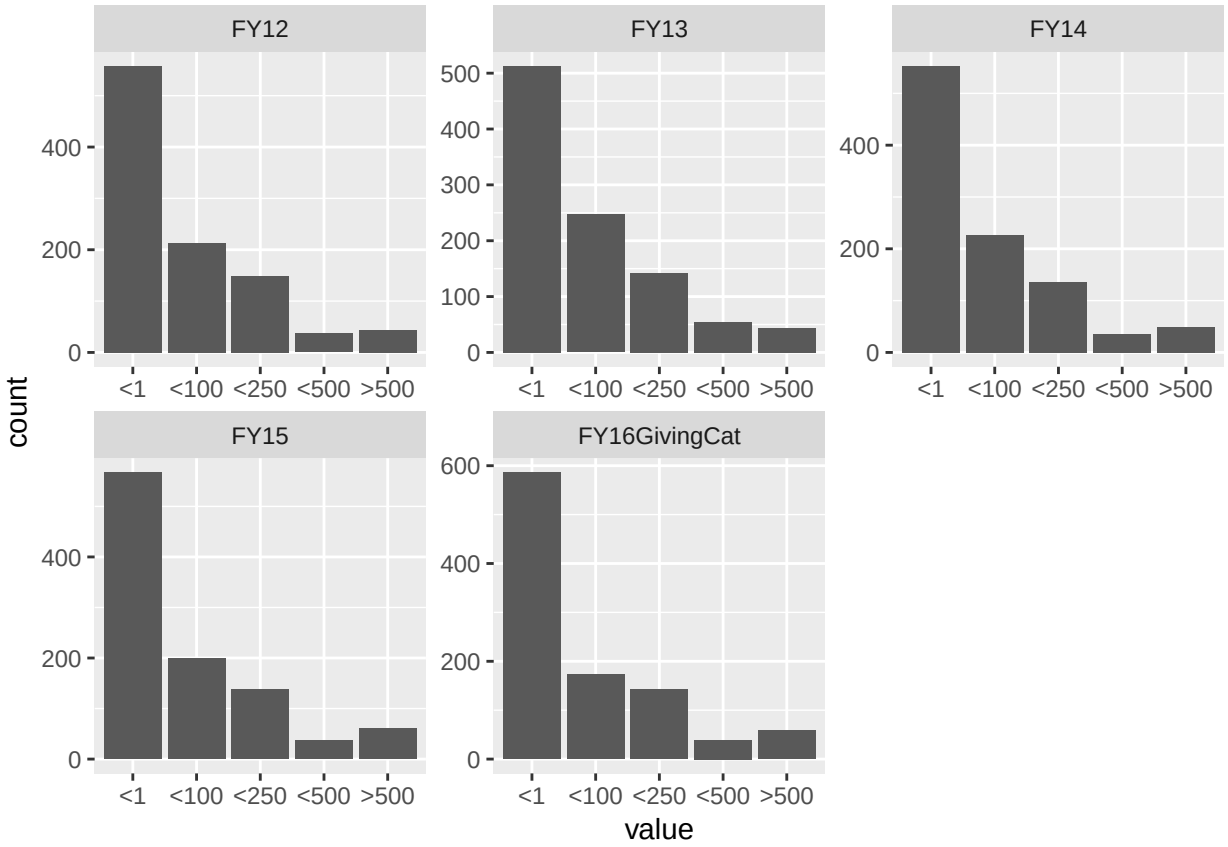
2.2.1 Univariate Analysis:

Contributions:

Let's examine each variable first starting with the "variable of interest" - 2016 contributions.

In 2016, roughly half the alumni population did not give anything. 20-25%% gave less than $100; Around 15% gave $100-250; About 4% gave $250-500 and less than 1% give more than $500. We see from the figures that the alumni contributions in all years followed similar distributions. In 2013, we had the highest percentage of the Alumni contributing (about 49%) and 2015 marked the lowest % giving (43%)

```
lcat <- c("FY16GivingCat", "FY14", "FY15", "FY13", "FY12")
dt[, lcat, with = FALSE] %>% na.omit() %>% gather() %>% ggplot(aes(value)) +
    facet_wrap(~key, scales = "free") + scale_x_discrete(labels = c("<1",
    "<100", "<250", "<500", ">500")) + geom_bar()
```



###Other

Categorical variables:

We'll take a look at the distribution of the other categorical variables

Gender:

```
prop.table(table(dt$Gender))
```

```
##
##     F     M
## 0.505 0.495
```

Gender is nearly evenly distributed between male and female.

Class Year:

```
prop.table(table(dt$Class.Year))
```

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

Only ten percent of alumni graduated in the 1970s and the largest proportion graduated around the 2010's. In fact, the the relationship is monotonic, where the younger the cohort, the larger the cohort.

Marital Status:

```
prop.table(table(dt$MS))
```

```
##
##     D     M     S     W
## 0.061 0.584 0.344 0.011
```

More than half are married, with only 6 percent divorced and 1 percent widowed

Major:

```
prop.table(table(dt$MajorCat))
```

```
##
##         HUM_ART           OTHER SOCIAL_SCIENCE            STEM
##           0.451           0.075           0.260           0.214
```

Humanities and Arts comprise almost half the majors, with social science second at more than a squarter and STEM at just over 20%

Next Degree:

```
prop.table(table(dt$NextDeg))
```

```
##
##     0     1
## 0.378 0.622
```

Mroe than 60% of alumni have a higher degree

2.2.2 Bivariate Analysis:

Giving in 2016 vs. 2015:

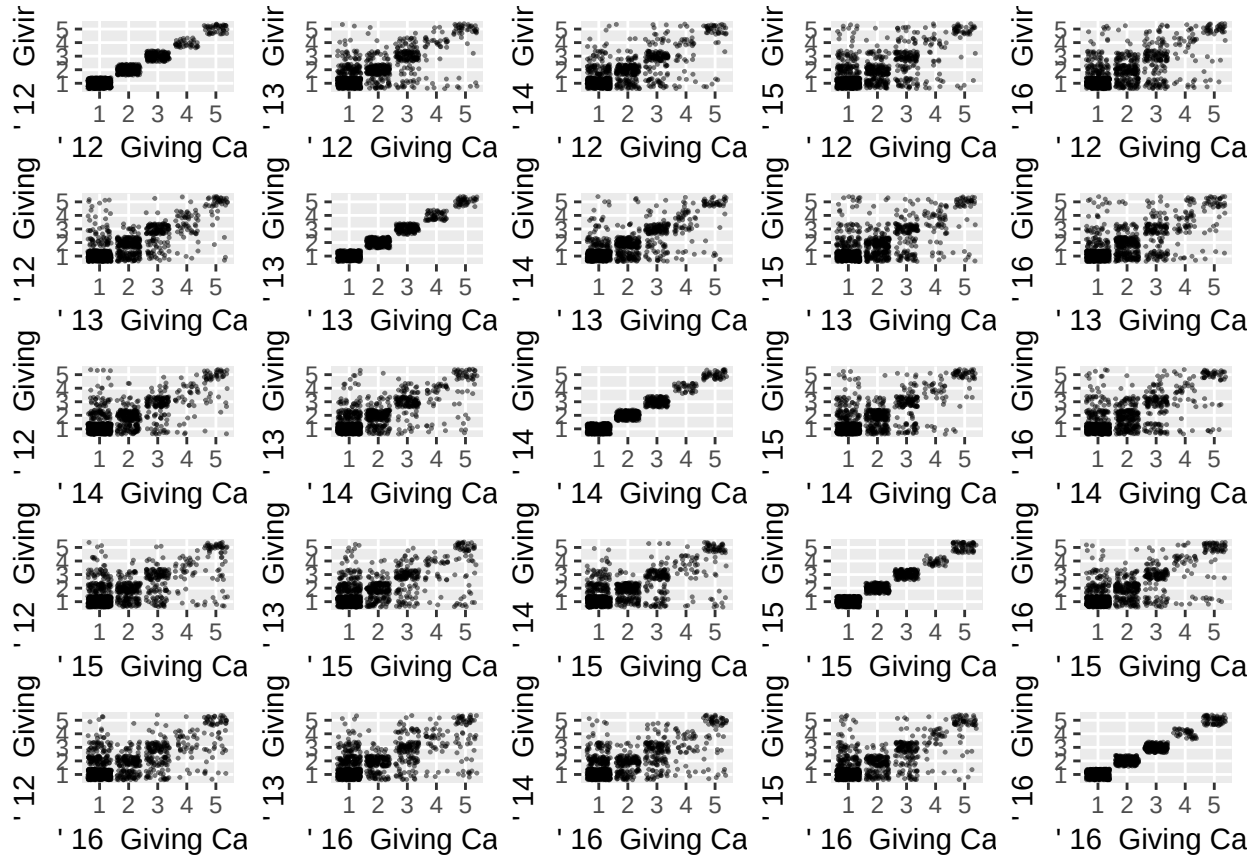
Let's look at how the giving categories relate to each other.

```
ggScatterJitter <- function(yearcat1, yearcat2) {
    if (yearcat1 == 16) {
        yeargive1 <- "FY16GivingCat"
    } else {
        yeargive1 <- paste("FY", toString(yearcat1), sep = "")
    }
    if (yearcat2 == 16) {
        yeargive2 <- "FY16GivingCat"
    } else {
        yeargive2 <- paste("FY", toString(yearcat2), sep = "")
    }

    p <- ggplot(dt, aes_string(x = yeargive1, y = yeargive2))
    p <- p + geom_jitter(size = 0.2, alpha = 0.5)
    p <- p + scale_x_discrete(name = paste("'", toString(yearcat1),
        " Giving Cat"), labels = c("1", "2", "3", "4", "5")) +
```

```
        scale_y_discrete(name = paste("'", toString(yearcat2),
            " Giving"), labels = c("1", "2", "3", "4", "5"))
    return(p)
}
plots <- list()
k <- 1
for (i in 12:16) {
    for (j in 12:16) {
        plots[[k]] <- ggScatterJitter(i, j)
        k = k + 1
    }
}

grid.arrange(grobs = plots, ncol = 5)
```



We observe that there is a high correlation between a person's giving one year and giving in another. Essentially, someone's giving category in 2016 is most likely to be the same as their past; it is also interesting to note that in most instances the second highest category is the "[0,1)" category - so basically either they give like they have given in the past or not give at all!

```
l = c("FY12Giving", "FY13Giving", "FY14Giving", "FY15Giving",
    "FY16Giving")
cor(dt[, l, with = FALSE])
```

```
##             FY12Giving FY13Giving FY14Giving FY15Giving FY16Giving
## FY12Giving   1.0000000  0.2503262  0.8764810  0.3842109  0.6887085
## FY13Giving   0.2503262  1.0000000  0.2077211  0.1376795  0.1007133
```

```
## FY14Giving  0.8764810  0.2077211  1.0000000  0.6019425  0.8238437
## FY15Giving  0.3842109  0.1376795  0.6019425  1.0000000  0.5899564
## FY16Giving  0.6887085  0.1007133  0.8238437  0.5899564  1.0000000
```

We can see that all the giving variables are correlated to each other. The strongest correlation is around .88 and the weakest .10. This informs us that a model with too many of these variables might be highly colinear and thus overfit.

```
GenXtab = function(dframe, x1, x2, nlist) {
    x1vsx2 = xtabs(formula = ~x1 + x2, data = dframe)
    names(dimnames(x1vsx2)) = nlist
    print(x1vsx2)
    print("Percentage of Column Totals Shown Below")
    print(round(prop.table(x1vsx2, 2), 2))
    a.s = assocstats(x1vsx2)
    a.s
    if (is.null(a.s$phi) | is.na(a.s$phi)) {
        phi_s = "Phi: N/A;"
    } else if (abs(a.s$phi) > 0.3) {
        phi_s = "Phi: Medium;"
    } else if (abs(a.s$phi) > 0.1) {
        phi_s = "Phi: Small;"
    } else {
        phi_s = "Phi: Negligible;"
    }

    if (is.null(a.s$contingency)) {
        cc_s = "Contingency Coef: N/A;"
    } else if (abs(a.s$contingency) > 0.5) {
        cc_s = "Contingency Coef: Large;"
    } else if (abs(a.s$contingency) > 0.3) {
        cc_s = "Contingency Coef: Medium;"
    } else if (abs(a.s$contingency) > 0.1) {
        cc_s = "Contingency Coef: Small;"
    } else {
        cc_s = "Contingency Coef: Negligible;"
    }

    if (is.null(a.s$cramer)) {
        c_s = "Cramer's V: N/A;"
    } else if (abs(a.s$cramer) > 0.5) {
        c_s = "Cramer's V: Large;"
    } else if (abs(a.s$cramer) > 0.3) {
        c_s = "Cramer's V: Medium;"
    } else if (abs(a.s$cramer) > 0.1) {
        c_s = "Cramer's V: Small ;"
    } else {
        c_s = "Cramer's V: Negligible;"
    }
    print(paste("Assoc Stats Conclusions: ", phi_s, cc_s, c_s))
}
```

Giving in 2016 vs. Alumni Event Attendance:

Attendance at an alumni event is the factor (other than previous years' contributions), that our intuitions syas will have the strongest relationship with 2016 give.

There is a relationship between attendence at alumni events in 2012-2015 and 2016 giving. 50% of those attending gave in 2016 while only 28% of those not attending gave in 2016.See figures below for reference

```
GenXtab(dframe = dt, x1 = dt$Event, x2 = dt$FY16GivingCat, nlist = c("Attendence",
    "FY16Cat"))
```

```
##           FY16Cat
## Attendence [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##          0   286      61        36         5           7
##          1   300     112       107        34          52
## [1] "Percentage of Column Totals Shown Below"
##           FY16Cat
## Attendence [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##          0  0.49    0.35      0.25      0.13        0.12
##          1  0.51    0.65      0.75      0.87        0.88
## [1] "Assoc Stats Conclusions:  Phi: N/A; Contingency Coef: Small; Cramer's V: Small ;"
```

Unfortunately, by contingency coefficient and Cramer's V, the relationship is a weak one.

Giving in 2016 by Marital status,:

Most of the Alumni fall into either Married or Single category and Married Alumni are more likely to give than Single.
There are very few data points for Widowed and Divorced alumni - so we do not want to make broad conclusions, but it appears that there are both Divorced and Widowed Alumni that contribute high amounts (and there are those that contribute nothing too in these categories) One possible reason we are seeing Married giving more than Single might actually have to do with age. Older Alumni are more likely to be married and older alumni are also probably a bit more well established financially - so more likely to contribute to charitable causes. So Marital status vs. Giving might simply be capturing the relationship between Age and Giving. While age is not a variable that's available in the dataset, we have "Class Year" which is a good proxy for age.

```
GenXtab(dframe = dt, x1 = dt$MS, x2 = dt$FY16GivingCat, nlist = c("Marital",
    "FY16Cat"))
```

```
##         FY16Cat
## Marital [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##       D    36       9        11         2           3
##       M   305      96       109        31          43
##       S   241      66        23         4          10
##       W     4       2         0         2           3
## [1] "Percentage of Column Totals Shown Below"
##         FY16Cat
## Marital [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##       D  0.06    0.05      0.08      0.05        0.05
##       M  0.52    0.55      0.76      0.79        0.73
##       S  0.41    0.38      0.16      0.10        0.17
```

```
##       W  0.01    0.01      0.00      0.05        0.05
## [1] "Assoc Stats Conclusions:  Phi: N/A; Contingency Coef: Small; Cramer's V: Small ;"
```

And again, we see that the the relationship between marital status and giving is small.

Giving 2016 vs. Class Year:

In the figure below, we filter out all the non-givers, and we see that older graduates tend to have donated more, given that they donated anything at all.

```
GenXtab(dframe = dt, x1 = dt$Class, x2 = dt$FY16GivingCat, nlist = c("Class",
    "FY16Cat"))
```

```
##       FY16Cat
## Class  [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##   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
## [1] "Percentage of Column Totals Shown Below"
##       FY16Cat
## Class  [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##   1972  0.09    0.05      0.16      0.18        0.27
##   1982  0.15    0.13      0.24      0.36        0.25
##   1992  0.20    0.17      0.27      0.23        0.20
##   2002  0.23    0.24      0.17      0.15        0.24
##   2012  0.33    0.42      0.15      0.08        0.03
## [1] "Assoc Stats Conclusions:  Phi: N/A; Contingency Coef: Small; Cramer's V: Small ;"
```

Again, we find that there is a weak relationship between class year and giving.

Giving in 2016 by Gender:

There is no statistical evidence that whether or not Alumni Give in 2016 varies by Gender (41% of Females Gave vs. 42% of Males). However when we look at the categories of contribution in 2016 by Gender, we see differences worth investigating later. There may be other factors at play here, for e.g. we know that older alumni tend to give more than younger alumni; it is possible that there are fewer "older" female alumni (fewer women attended college in 1972) than Male.

```
GenXtab(dframe = dt, x1 = dt$Gender, x2 = dt$FY16GivingCat, nlist = c("Gender",
    "FY16Cat"))
```

```
##       FY16Cat
## Gender [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##      F   298     106        58        17          26
##      M   288      67        85        22          33
## [1] "Percentage of Column Totals Shown Below"
##       FY16Cat
## Gender [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##      F  0.51    0.61      0.41      0.44        0.44
##      M  0.49    0.39      0.59      0.56        0.56
```

```
## [1] "Assoc Stats Conclusions:  Phi: N/A; Contingency Coef: Small; Cramer's V: Small ;"
```

Again, we see that the relation between gender and giving is weak.

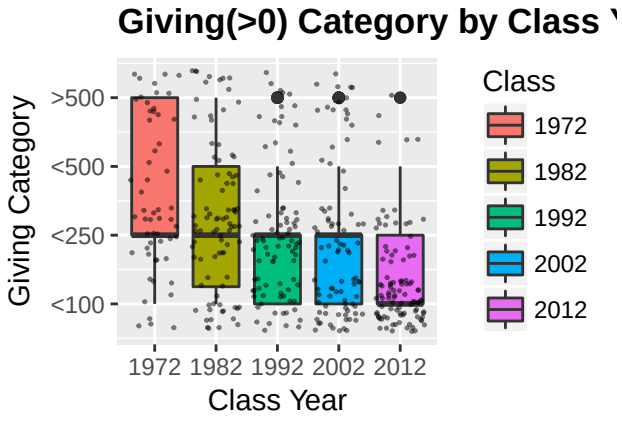
```
GenXtab(dframe = dt, x1 = dt$Gender, x2 = dt$Class, nlist = c("Gender",
    "Class"))
```

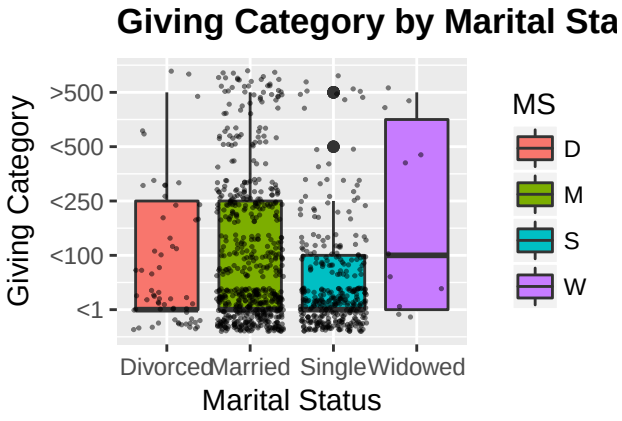
```
##       Class
## Gender 1972 1982 1992 2002 2012
##      F   38   80  102  133  152
##      M   67   96  101   90  141
## [1] "Percentage of Column Totals Shown Below"
##       Class
## Gender 1972 1982 1992 2002 2012
##      F 0.36 0.45 0.50 0.60 0.52
##      M 0.64 0.55 0.50 0.40 0.48
## [1] "Assoc Stats Conclusions:  Phi: N/A; Contingency Coef: Small; Cramer's V: Small ;"
```

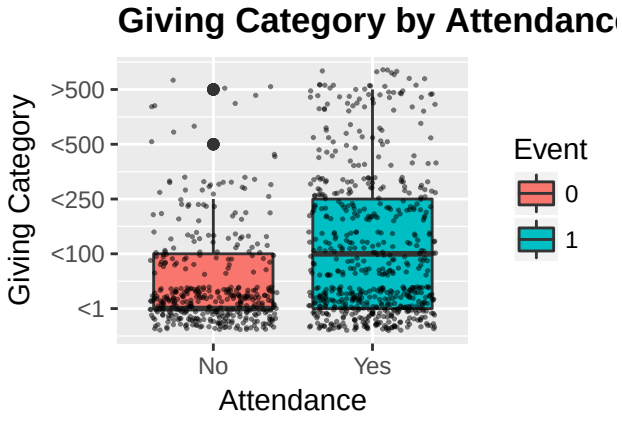
Figures:

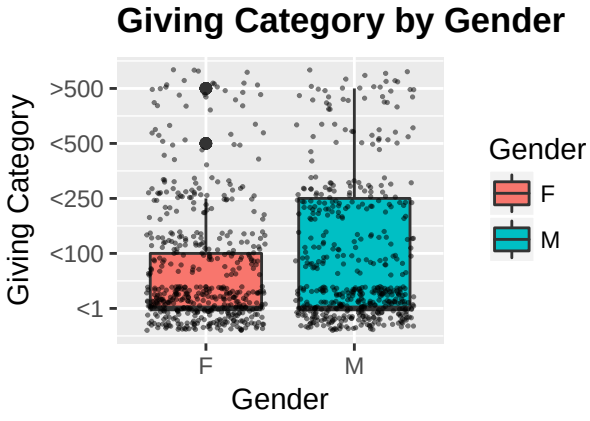
```
Year <- ggplot(dt[dt$FY16Giving > 0, ], aes(Class, as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = Class)) + ggtitle("Giving(>0) Category by Class Year") +
    geom_jitter(size = 0.25, alpha = 0.5) + scale_x_discrete(name = "Class Year") +
    scale_y_continuous(name = "Giving Category", breaks = 1:5,
        labels = c("<1", "<100", "<250", "<500", ">500")) + theme(plot.title = element_text(lineh
    face = "bold"))
Marriage <- ggplot(dt, aes(MS, as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = MS)) + ggtitle("Giving Category by Marital Status") +
    geom_jitter(size = 0.25, alpha = 0.5) + scale_x_discrete(name = "Marital Status",
    labels = c("Divorced", "Married", "Single", "Widowed")) +
    scale_y_continuous(name = "Giving Category", breaks = 1:5,
        labels = c("<1", "<100", "<250", "<500", ">500")) + theme(plot.title = element_text(lineh
    face = "bold"))
Attendence <- ggplot(dt, aes(factor(Event), as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = factor(Event))) + ggtitle("Giving Category by Attendance at Event") +
    geom_jitter(size = 0.25, alpha = 0.5) + scale_x_discrete(name = "Attendance",
    labels = c("No", "Yes")) + scale_y_continuous(name = "Giving Category",
    breaks = 1:5, labels = c("<1", "<100", "<250", "<500", ">500")) +
    scale_fill_discrete(name = "Event") + theme(plot.title = element_text(lineheight = 1,
    face = "bold"))
Gender <- ggplot(dt, aes(Gender, as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = Gender)) + ggtitle("Giving Category by Gender") +
    geom_jitter(size = 0.25, alpha = 0.5) + scale_x_discrete(name = "Gender") +
    scale_y_continuous(name = "Giving Category", breaks = 1:5,
        labels = c("<1", "<100", "<250", "<500", ">500")) + theme(plot.title = element_text(lineh
    face = "bold"))

grid.arrange(Year, Marriage, Attendence, Gender, ncol = 2)
```









Bivariate Analysis Conclusion:

Analysis of the variables shows that the most significant predictor of FY16 giving is likely going to be the giving from a previous year. Adding all the other year's giving might help our model, or it might cause issues with colinearity. We can add other covariates, such as class year or attendance at an alumni event, but they are unlikely to add much.

```
GenXtab(dframe = dt, x1 = dt$FY15, x2 = dt$FY16GivingCat, nlist = c("FY16",
    "FY14"))
```

```
##              FY14
## FY16          [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##   [0,1)         480      57        23         3           4
##   [1,100)        64     108        25         1           1
##   [100,250)      29       8        88        10           3
##   [250,500)       5       0         4        23           4
##   [500,2e+05)     8       0         3         2          47
## [1] "Percentage of Column Totals Shown Below"
##              FY14
## FY16          [0,1) [1,100) [100,250) [250,500) [500,2e+05)
##   [0,1)        0.82    0.33      0.16      0.08        0.07
##   [1,100)      0.11    0.62      0.17      0.03        0.02
##   [100,250)    0.05    0.05      0.62      0.26        0.05
##   [250,500)    0.01    0.00      0.03      0.59        0.07
##   [500,2e+05)  0.01    0.00      0.02      0.05        0.80
## [1] "Assoc Stats Conclusions:  Phi: N/A; Contingency Coef: Large; Cramer's V: Large;"
```