Lab 02

w271

Alex Yang, John Kenney, Ram Balasubramanian

Oct 12, 2017

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 demographic information (like gender, marital status etc.), some university specific information (graduation year, major of studies), and some post-graduation information (Alumni event attendance, historical contributions).
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.

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

#Libraries required
library(car)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:car':
##
##     recode
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(Hmisc)#Used by author for 3D plotting
```

```
## 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:dplyr':
##
##     combine, src, summarize
```

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

```
library(ggplot2)
library(gridExtra)
```

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

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

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

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

```
library(effsize) #Used to calculate Cohen's D for T-Test
```

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

```
library(aod)    #Used for effect size of the logit model
```

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

```
##
## Attaching package: 'aod'
```

```
## The following object is masked from 'package:survival':
##
##     rats
```

```
library(mcprofile) #Used for confidence intervals

library(data.table)
```

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

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

Read Data

```
dt <- fread("lab2data.csv")
describe(dt)
```

```
## dt
##
##  12  Variables      1000  Observations
## ---------------------------------------------------------------------------
## V1
##        n  missing distinct
##     1000        0     1000
##
## lowest : 1    10   100  1002 1003, highest: 995  996  997  998  999
## ---------------------------------------------------------------------------
## 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             Ch
## highest: Spanish              Speech (Drama, etc.) Speech Correction    Theatre             Zo
## ---------------------------------------------------------------------------
## 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
## ---------------------------------------------------------------------------
```

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.

Exploratory Data Analysis:

Univariate Analysis:

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

FY16Giving: Given that the vast majority of people did not give in 2016 and the skewness of the data (with a few large contributions) - let's also look at the distribution after a log-transformation (this is something we may want to consider for our modeling purposes)

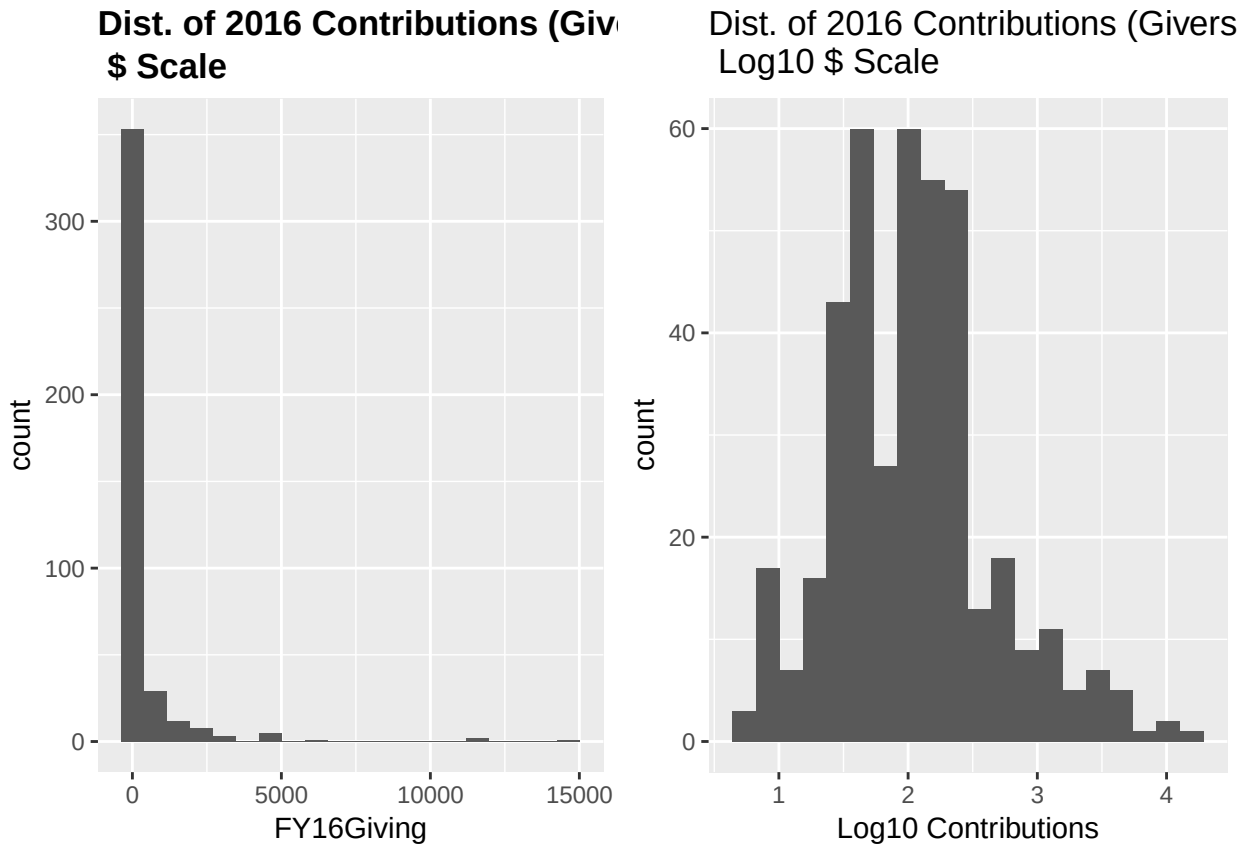
```
h1 = ggplot(data = dt[FY16Giving > 0], aes(x = FY16Giving)) +
    geom_histogram(bins = 20) + ggtitle("Dist. of 2016 Contributions (Givers Only) \n $ Scale") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))

h2 = ggplot(data = dt[FY16Giving > 0], aes(x = log10(FY16Giving))) +
    geom_histogram(bins = 20) + ggtitle("Dist. of 2016 Contributions (Givers Only) \n Log10 $ Scal
    xlab("Log10 Contributions")
theme(plot.title = element_text(lineheight = 1, face = "bold"))
```

```
## List of 1
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : chr "bold"
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : num 1
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
```

```
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
grid.arrange(h1, h2, ncol = 2)
```

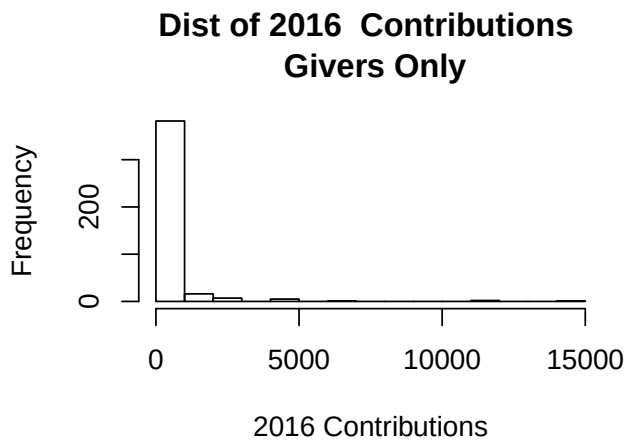


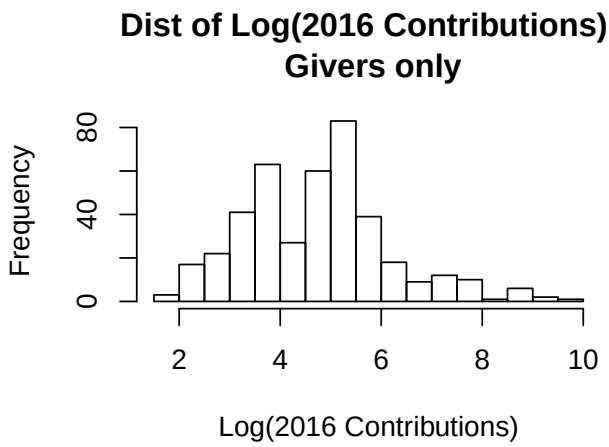
The distribution of log wages is somewhat normal, with the majority less than $100, and a tail that stretches into 4 digits

Log is base 10 because meaning is more inuitive than a natural log

"' Let's look at a distribution of just the "givers" (i.e. take out the zero contributions) to get a better picture.

```
par(mfrow = c(1, 2))
hist(dt[FY16Giving > 0]$FY16Giving, breaks = 20, main = "Dist of 2016  Contributions \n Givers Onl
    xlab = "2016 Contributions")
hist(log(dt[FY16Giving > 0]$FY16Giving), breaks = 20, main = "Dist of Log(2016 Contributions) \n (
    xlab = "Log(2016 Contributions)")
```





Let's group the 2016 contributions by the categories that the university is interested in.

```
dt$FY16GivingCat <- cut(dt$FY16Giving, c(0, 1, 100, 250, 500,
    2e+05), right = FALSE)
summary(dt$FY16GivingCat)
```

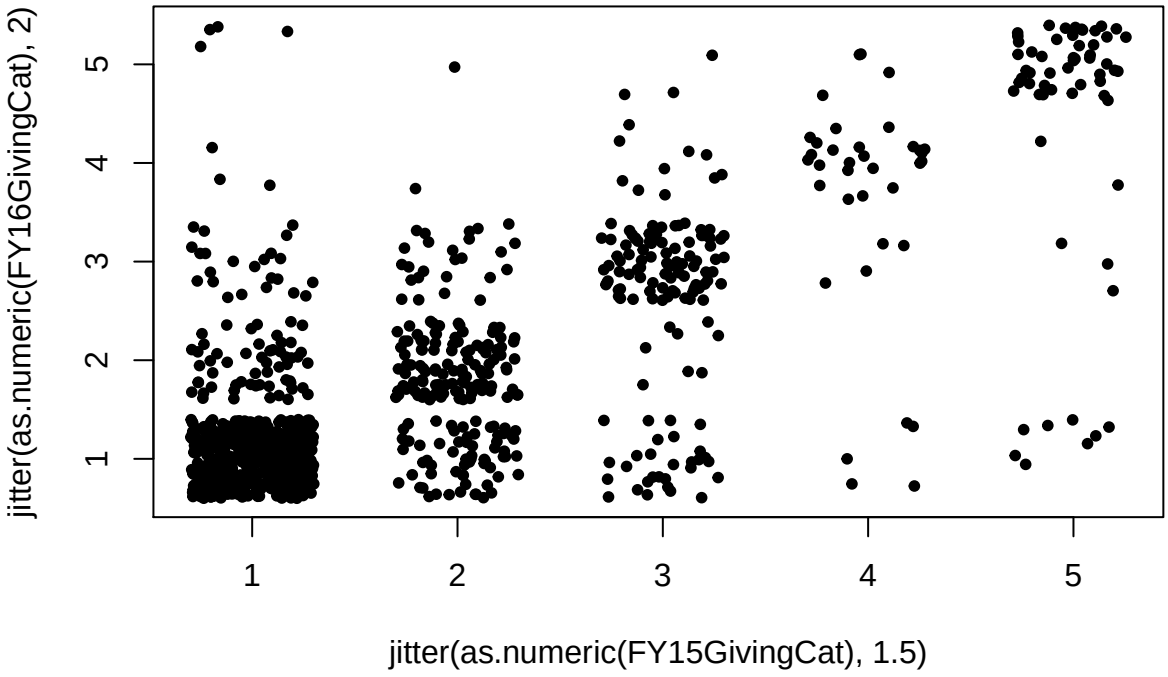
```
##       [0,1)     [1,100)   [100,250)   [250,500) [500,2e+05)
##         586         173         143          39          59
```

Do it for other years as well

```
dt$FY15GivingCat <- cut(dt$FY15Giving, c(0, 1, 100, 250, 500,
    2e+05), right = FALSE)
dt$FY14GivingCat <- cut(dt$FY14Giving, c(0, 1, 100, 250, 500,
    2e+05), right = FALSE)
dt$FY13GivingCat <- cut(dt$FY13Giving, c(0, 1, 100, 250, 500,
    2e+05), right = FALSE)
dt$FY12GivingCat <- cut(dt$FY12Giving, c(0, 1, 100, 250, 500,
    2e+05), right = FALSE)
```

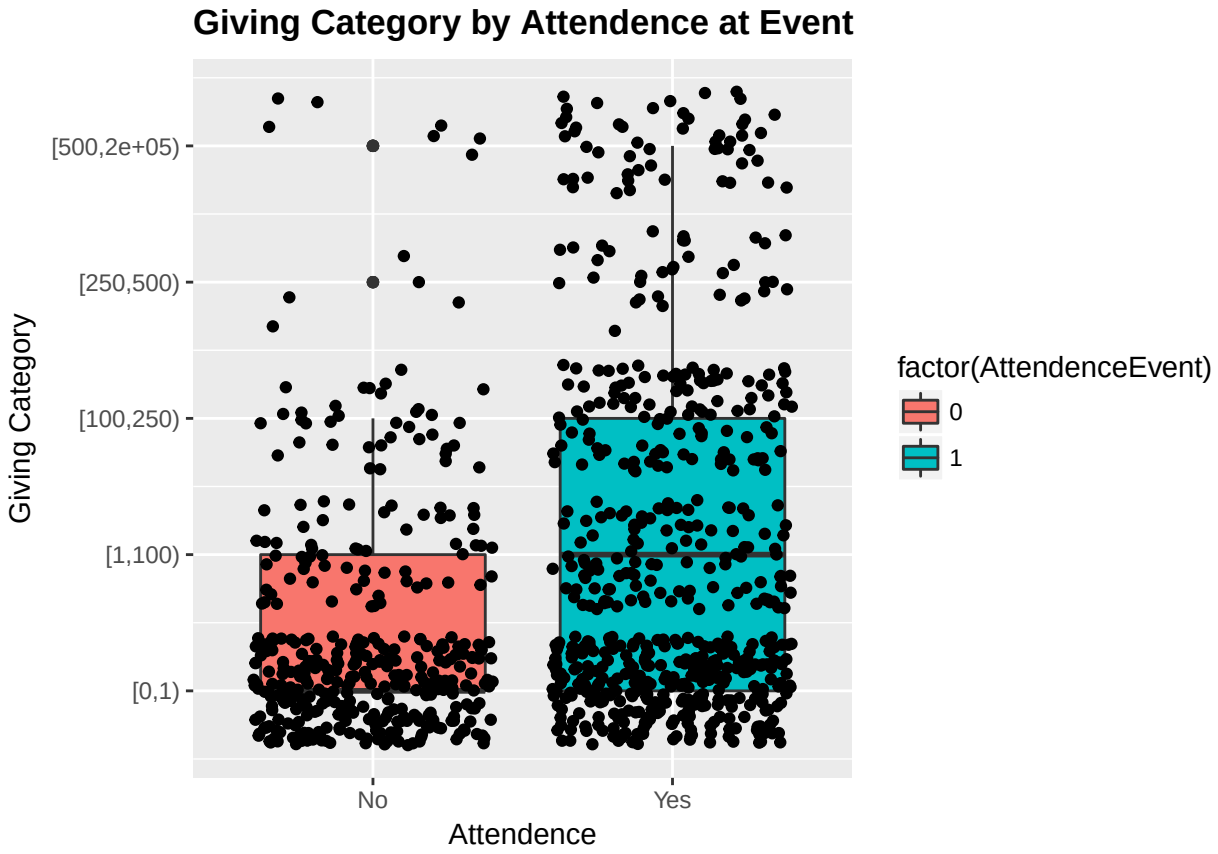
And do a scatterplot matrix

```
plot(jitter(as.numeric(FY16GivingCat), 2) ~ jitter(as.numeric(FY15GivingCat),
    1.5), data = dt, pch = 20)
```



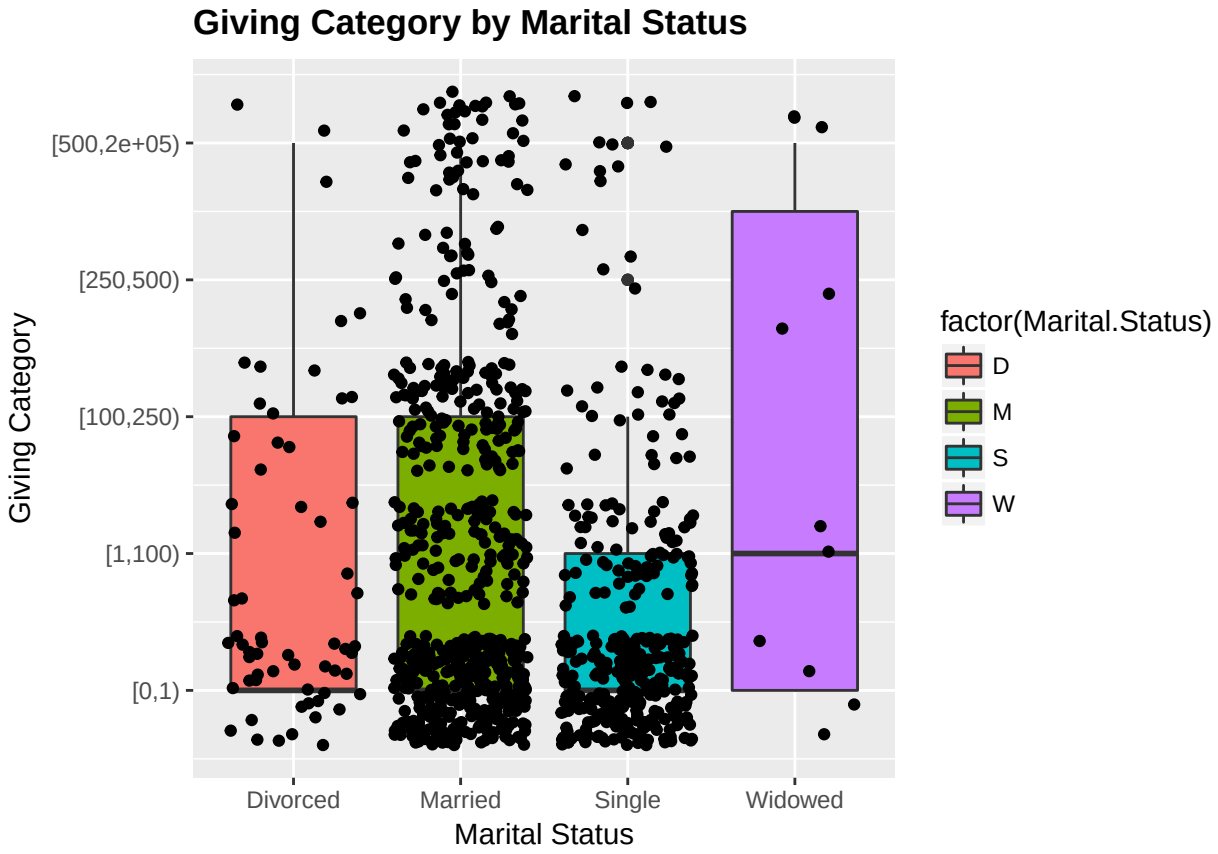
So there is strong correlation between giving in one year and another Distribution of log give (minus the 0's)

```
ggplot(dt, aes(factor(AttendenceEvent), as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = factor(AttendenceEvent))) + ggtitle("Giving Category by Attendence at
    geom_jitter() + scale_x_discrete(name = "Attendence", labels = c("No",
    "Yes")) + scale_y_continuous(name = "Giving Category", breaks = 1:5,
    labels = c("[0,1)", "[1,100)", "[100,250)", "[250,500)",
        "[500,2e+05)")) + theme(plot.title = element_text(lineheight = 1,
    face = "bold"))
```



So there does seem to be a relationship between giving in 2016 and attendence at the alumni event Try giving Category by marital status

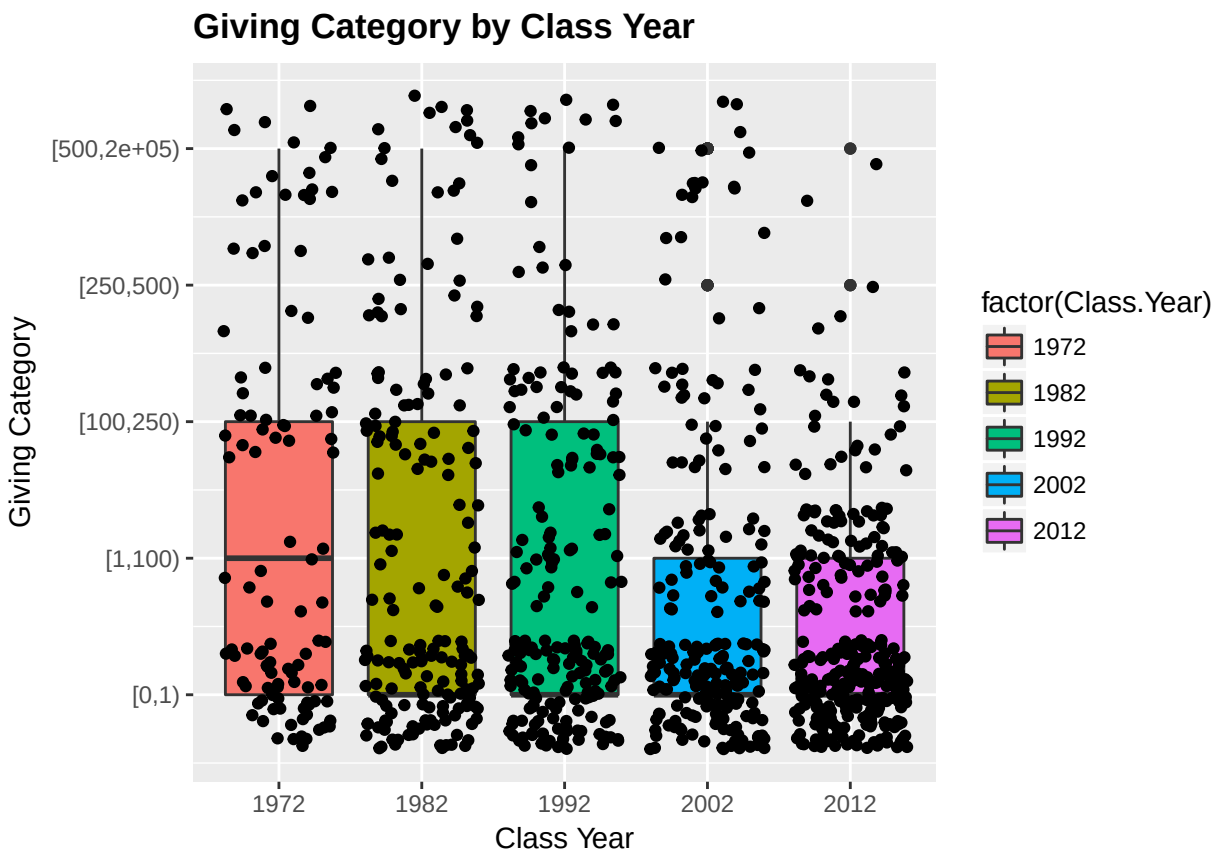
```
ggplot(dt, aes(factor(Marital.Status), as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = factor(Marital.Status))) + ggtitle("Giving Category by Marital Status'
    geom_jitter() + scale_x_discrete(name = "Marital Status",
    labels = c("Divorced", "Married", "Single", "Widowed")) +
    scale_y_continuous(name = "Giving Category", breaks = 1:5,
        labels = c("[0,1)", "[1,100)", "[100,250)", "[250,500)",
            "[500,2e+05)")) + theme(plot.title = element_text(lineheight = 1,
    face = "bold"))
```



there are too few data points for divorced and widowed, but it does seem that married people donate more than single people

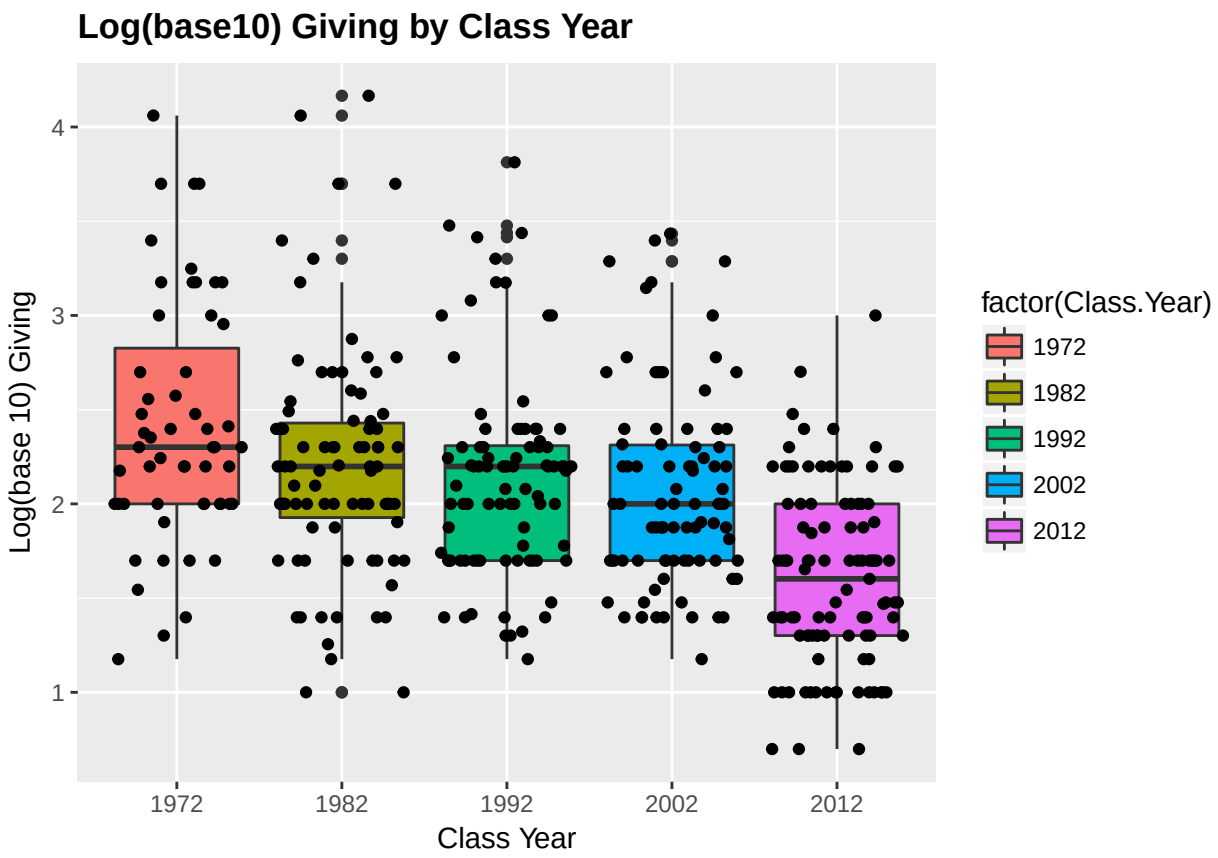
On the other hand, the distriubtion seems more even for widowed and divorced- there are just as many widowed alumni who give large amounts as there are who give little to nothing. This implies that giving might be related to age. Married people, after all, tend to be older than single people. so Marital status might just be capuring the effect of age or class Year

```
ggplot(dt, aes(factor(Class.Year), as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = factor(Class.Year))) + ggtitle("Giving Category by Class Year") +
    geom_jitter() + scale_x_discrete(name = "Class Year") + scale_y_continuous(name = "Giving Cate
    breaks = 1:5, labels = c("[0,1)", "[1,100)", "[100,250)",
        "[250,500)", "[500,2e+05)")) + theme(plot.title = element_text(lineheight = 1,
    face = "bold"))
```



Here, it indeed looks like the older the alumnus, the more likely he or she is to give money.Unfortunately, the boxplot by category misses a lot of data, so we can try to instead do a boxplot of the log of 2016 giving, minus the people who donated nothing

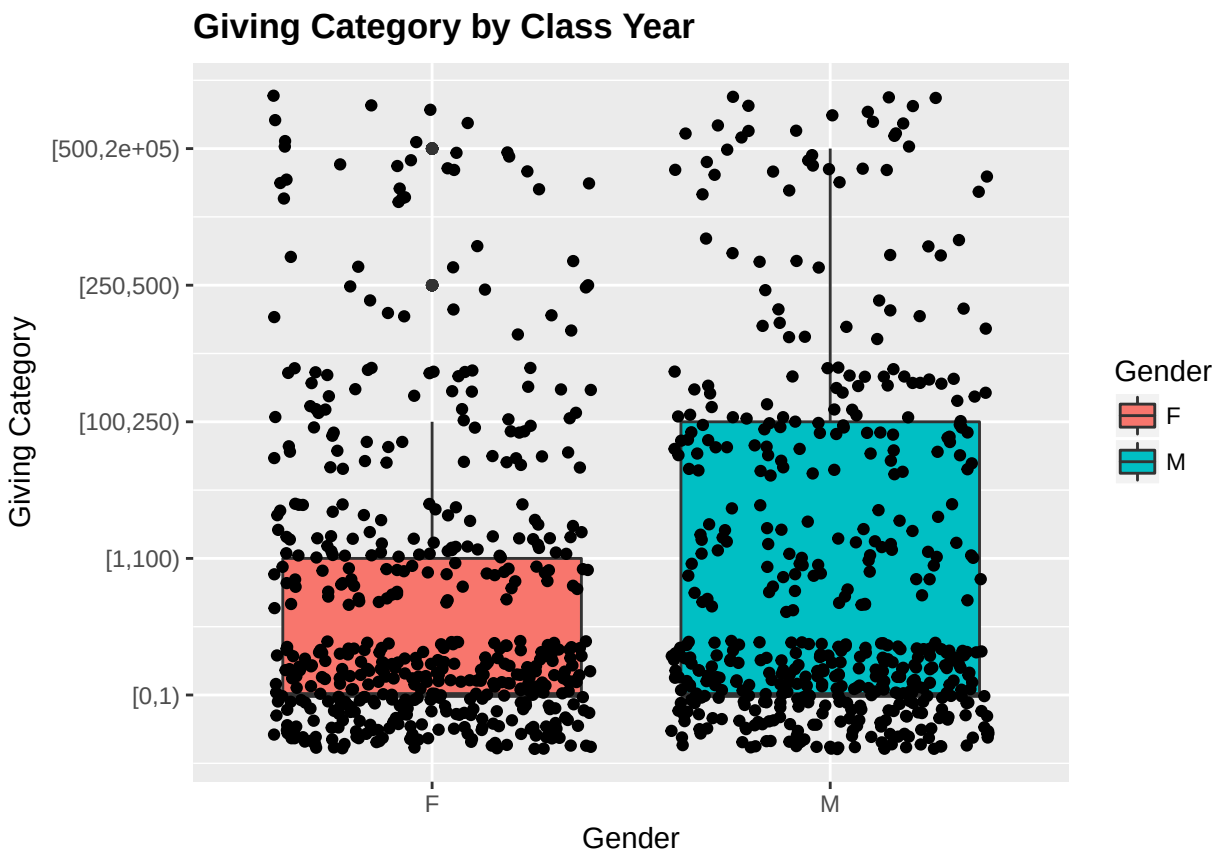
```
ggplot(dt[FY16Giving > 0], aes(factor(Class.Year), log10(FY16Giving))) +
    geom_boxplot(aes(fill = factor(Class.Year))) + ggtitle("Log(base10) Giving by Class Year") +
    geom_jitter() + scale_x_discrete(name = "Class Year") + scale_y_continuous(name = "Log(base 1
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



So for those who do donate, there seems to a clear trend that onlder alumni donate more

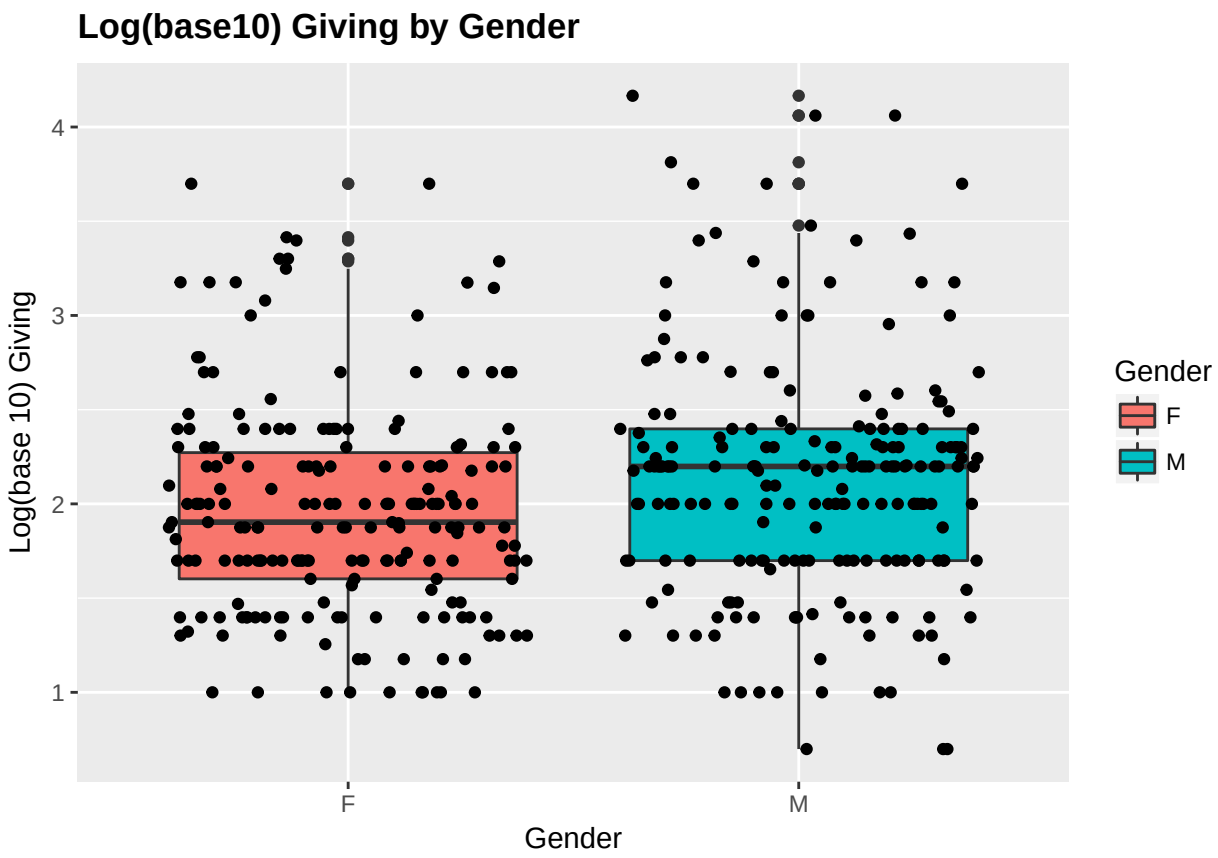
Is there a relationship in terms of gender?

```
ggplot(dt, aes(Gender, as.numeric(FY16GivingCat))) + geom_boxplot(aes(fill = Gender)) +
    ggtitle("Giving Category by Class Year") + geom_jitter() +
    scale_x_discrete(name = "Gender") + scale_y_continuous(name = "Giving Category",
    breaks = 1:5, labels = c("[0,1)", "[1,100)", "[100,250)",
        "[250,500)", "[500,2e+05)")) + theme(plot.title = element_text(lineheight = 1,
    face = "bold"))
```



It's a little hard to tell- since it does look like men donate a bit more to the university, but it's also possible that there are fewer female alumni (fewer women attended college in 1972)

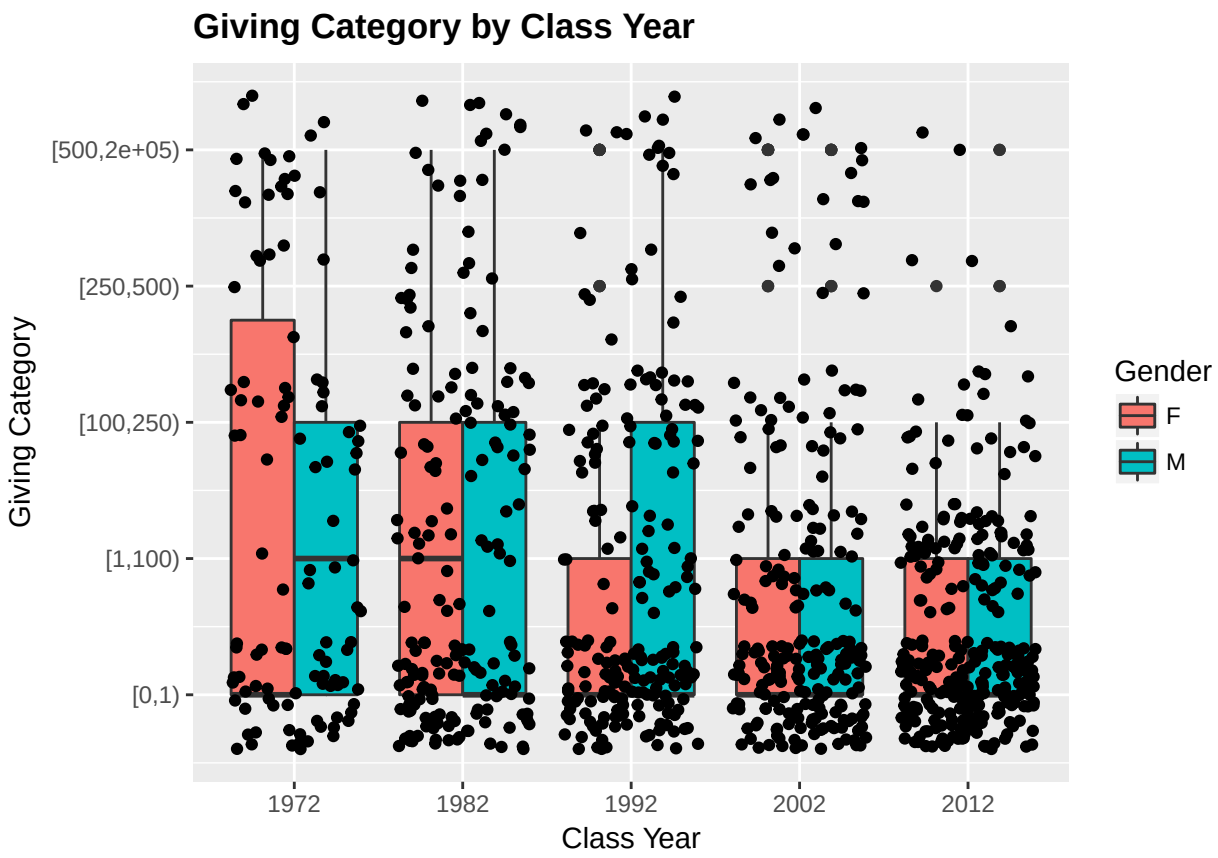
```
ggplot(dt[FY16Giving > 0], aes(Gender, log10(FY16Giving))) +
    geom_boxplot(aes(fill = Gender)) + ggtitle("Log(base10) Giving by Gender") +
    geom_jitter() + scale_x_discrete(name = "Gender") + scale_y_continuous(name = "Log(base 10) Gi
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



Of those that donate, men donate more. But we can't be sure this is significant as the median man donated less than the 75th percentile woman.

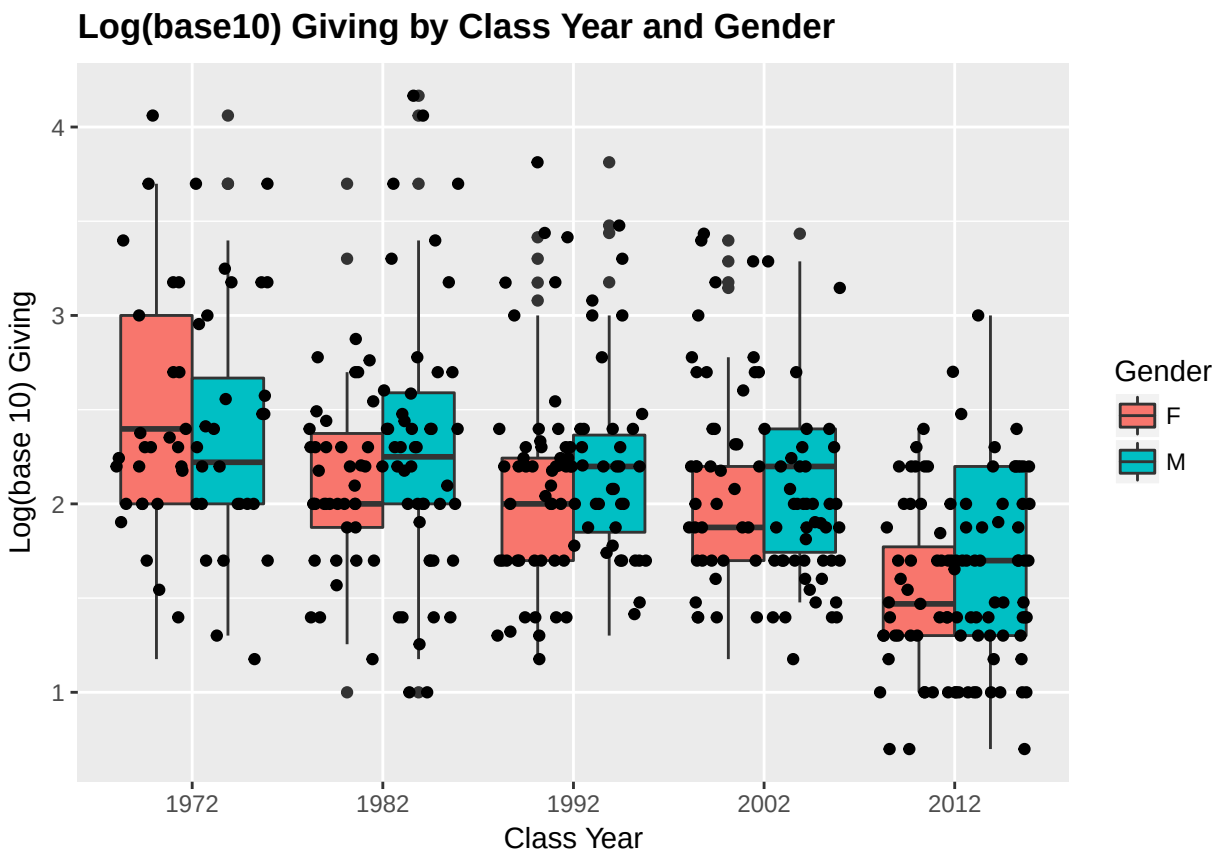
Try class year split by male and female:

```
ggplot(dt, aes(factor(Class.Year), as.numeric(FY16GivingCat))) +
    geom_boxplot(aes(fill = Gender)) + ggtitle("Giving Category by Class Year") +
    geom_jitter() + scale_x_discrete(name = "Class Year") + scale_y_continuous(name = "Giving Cate
    breaks = 1:5, labels = c("[0,1)", "[1,100)", "[100,250)",
        "[250,500)", "[500,2e+05)")) + theme(plot.title = element_text(lineheight = 1,
    face = "bold"))
```



So split by class year, gender does not seem to be a significant factor, and in fact older women donate more than older men

```
ggplot(dt[FY16Giving > 0], aes(factor(Class.Year), log10(FY16Giving))) +
    geom_boxplot(aes(fill = Gender)) + ggtitle("Log(base10) Giving by Class Year and Gender") +
    geom_jitter() + scale_x_discrete(name = "Class Year") + scale_y_continuous(name = "Log(base 1
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



SO limiting it to just alumni who donated in 2016, we do see that even split by age, men donated more than women, except for the oldest alumni