DataScience Capstone Project

Missy Lee
11/19/2016

Introduction

After recently closing our second fundraising campaign, we need to identify new potential major gift donors ahead of the next campaign. Major gifts are sought, not only to provide funding for buildings, scholarships, professorships and other initiatives but also attract other gifts. A major gift donor is currently defined as one who has the capacity to make a $25,000 gift. How can what we know about people who gave before be used to identify those who may give in the future? What characteristics can be identified and used to find currently unknown prospects? Do they currently give in smaller amounts (less than $25,000)? And if that is the case, is there a factor that converts someone from giving smaller amounts to giving more?

We can only try to answer these questions using the data already collected in the donor database. This study will explore a few variables to see how they might be used to model donor giving. The variables will be discussed in turn. The goal of this project is to examine how these selected variables influence lifetime giving among this group of donors and to evaluate their use in a regression model.

The dataset for this study is comprised of randomly sampled from a large database of known and potential donors. The variables chosen for this study were those easy to extract from the donor database without too much manipulation and have been shown in other studies to have some relationship to lifetime giving.

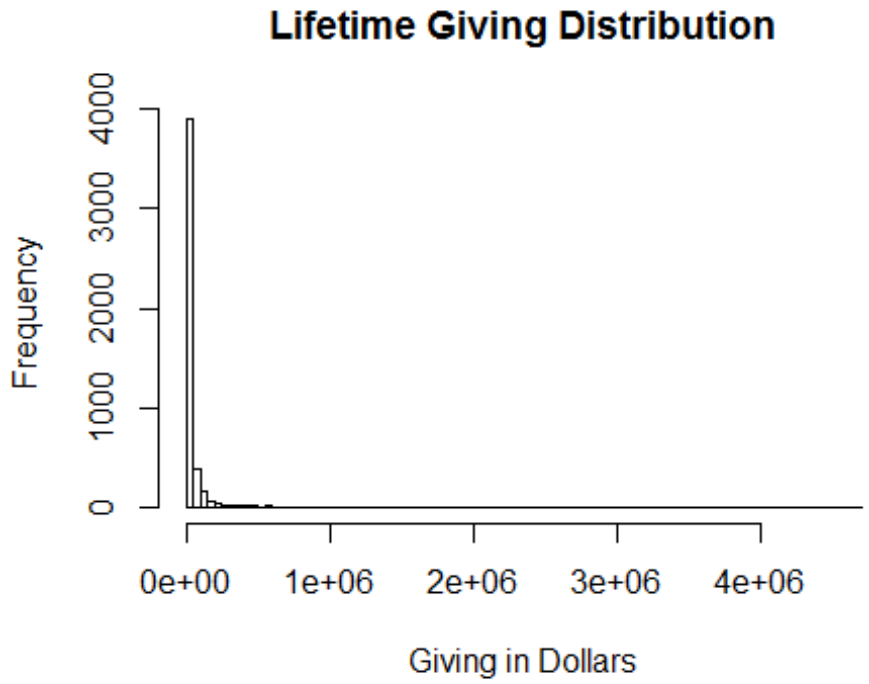
```
TS <- read.csv("~/GitHub/Mlee-Data-Science-Capstone-Project/TS.csv", header=T
RUE)
str (TS)
```

```
## 'data.frame':    7317 obs. of  15 variables:
##  $ IDCode                 : int  404 3198 3592 3085 7059 112 4832 3245 361
9 6147 ...
##  $ WSU.LIFETIME.GIVING    : num  17712 19475 6080 24155 0 ...
##  $ WSU.YEARS.OF.GIVING    : int  29 20 25 22 0 21 28 18 31 2 ...
##  $ ASSETS                 : int  8561566 7962000 7121842 5500000 4984000 3
205000 3137095 3064750 3054500 3029500 ...
##  $ RECORD.TYPE.CODE       : Factor w/ 13 levels "","AL","FA","FD",..: 2 2
2 2 4 2 2 2 2 2 ...
##  $ Number.of.relationships: int  1 1 3 9 NA 2 1 1 1 1 ...
##  $ Gender                 : Factor w/ 4 levels "","F","M","U": 3 3 3 3 3 3
##  $ Velocity35Score        : int  75 0 0 0 0 85 100 75 33 0 ...
##  $ Velocity57Score        : int  100 0 0 0 0 81 30 67 38 0 ...
##  $ AlumniCode             : int  1 1 1 1 0 1 1 1 1 1 ...
##  $ SportCode              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ GreekCode              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ParticipationScore     : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ AssetClass             : Factor w/ 6 levels "","Highest","Low",..: 6 6
##  $ ParticipationClass     : Factor w/ 5 levels "","Both","Greek",..: 2 2 2
2 2 2 2 2 2 2 ...
```

Lifetime Giving

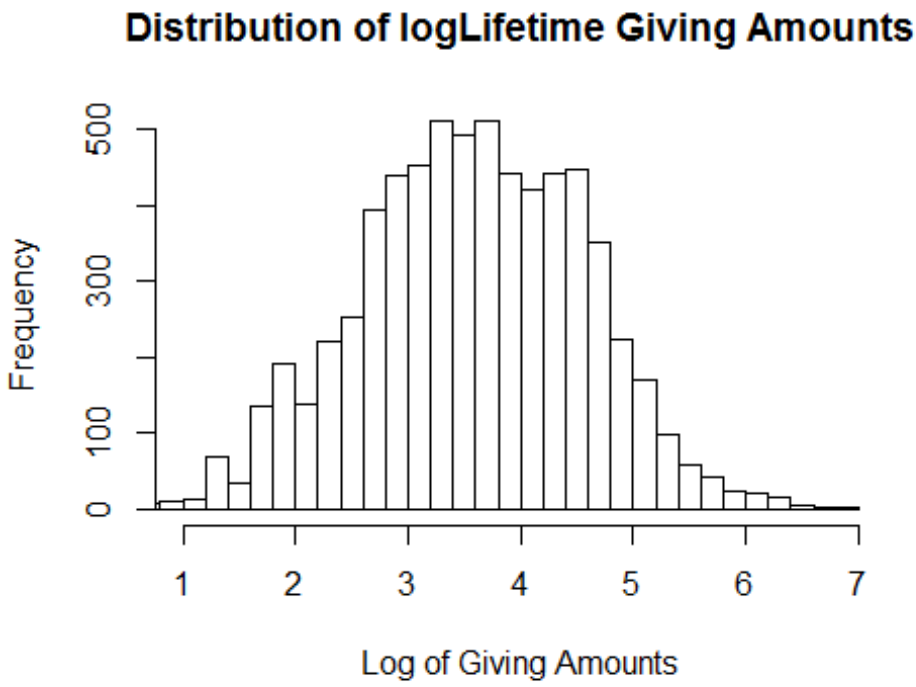
Lifetime Giving will be used as the dependent variable in this study. The values represented by this variable range from $0 to over $9,000,000. This range makes creating a histogram to show the distribution unweildy.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     450    2800   38280   17180 9673000       1
```

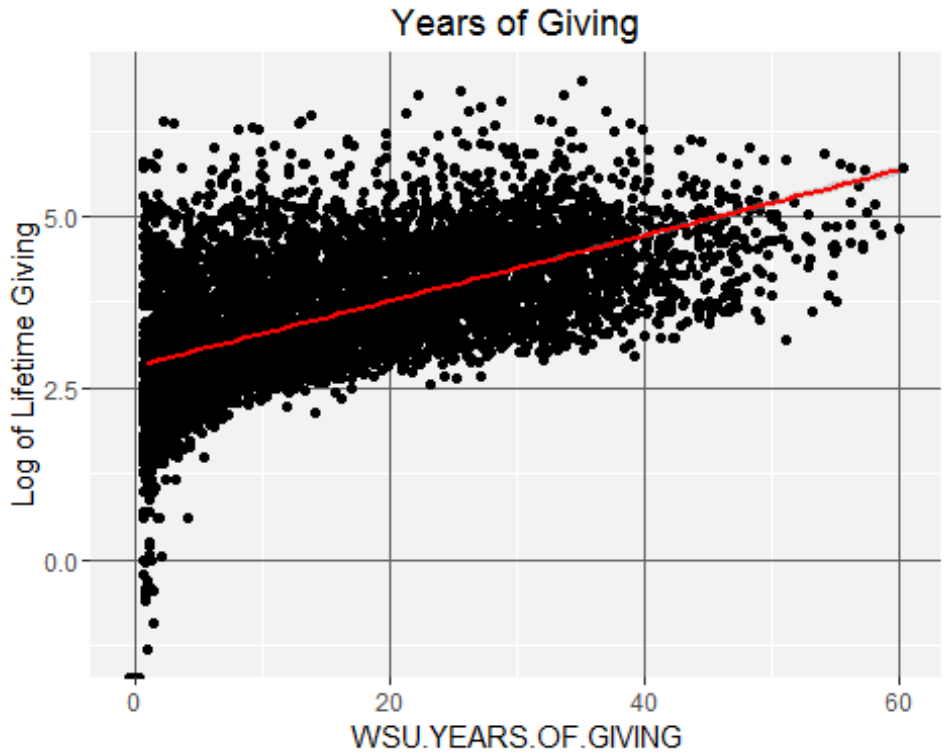


Many of the methods I intend to use require that the data be normally distributed. By transforming this data using a log10 transform, the data takes on a normal distribution. This transformed data will be easier to visualize since the data points will now be spread more uniformly in graphs. It will also be possible to capture linear relationships using linear regression with this transformed data.

```
 TS$logGiving <- log10(TS$WSU.LIFETIME.GIVING)
summary(TS$logGiving)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    -Inf   2.653   3.447    -Inf   4.235   6.986       1
```



Years of Giving



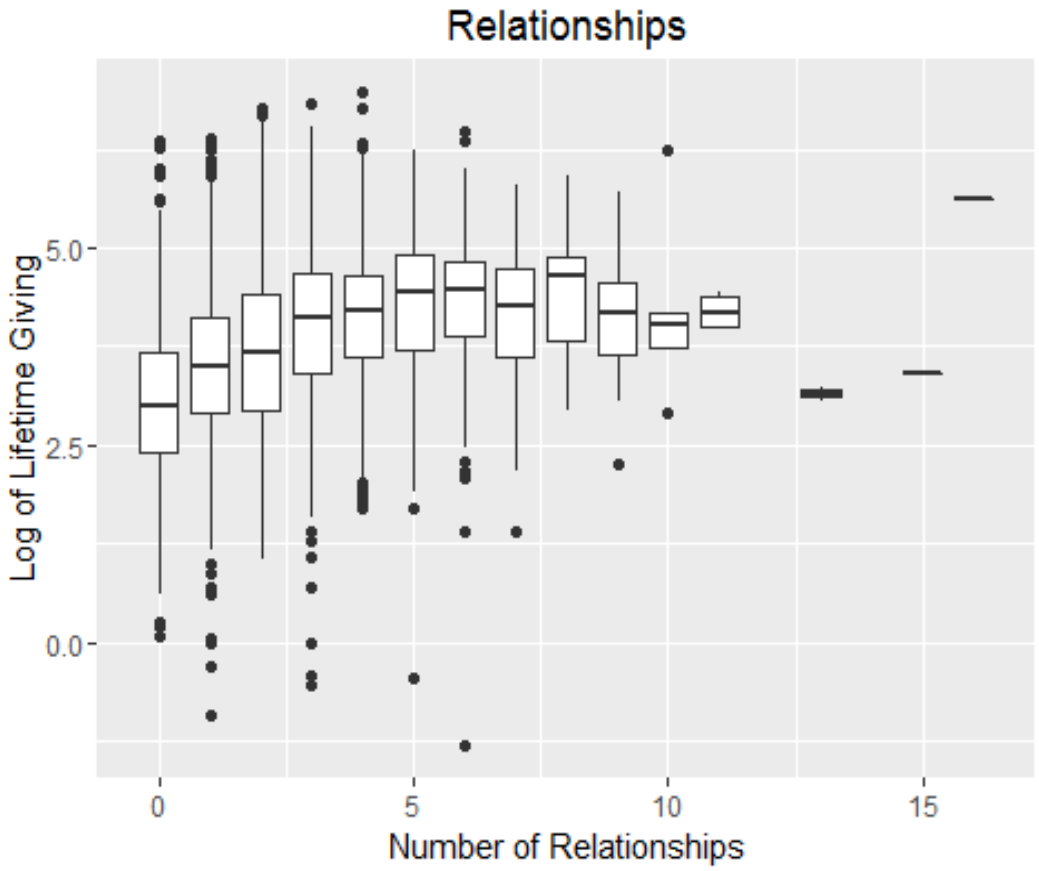
It is to be expected that giving over long amounts of time adds up to a large amount over a lifetime, and this plot does show that positive relationship. But there are enough points well above the line to say that many years of giving is not likely to be a requirement for large gifts.

The variable WSU.YEARS.OF.GIVING will be not included in the regression model.

A recommendation for future research is that this variable be examined in more detail. It is not necessarily true that the longer you have given, the more you have given. Size of the gifts through the years would make a great difference in cumulative giving, for example.

Number of relationships

Relationships in the donor database include parents, grandparents, aunts, uncles, siblings, as well as children who have also attended school here.



This plot seems to indicate a correlation between lifetime giving and a low to moderate number of other family members who attended the university.

The variable Number.of.Relationships will be included in the regression model since it seems to have a relationship to giving based Kendall's rank correlation. Kendall's correlation was used here because it deals better with ties than Spearman's correlation. Attempting to use Spearman resulted in an error "Cannot compute exact p-value with ties".

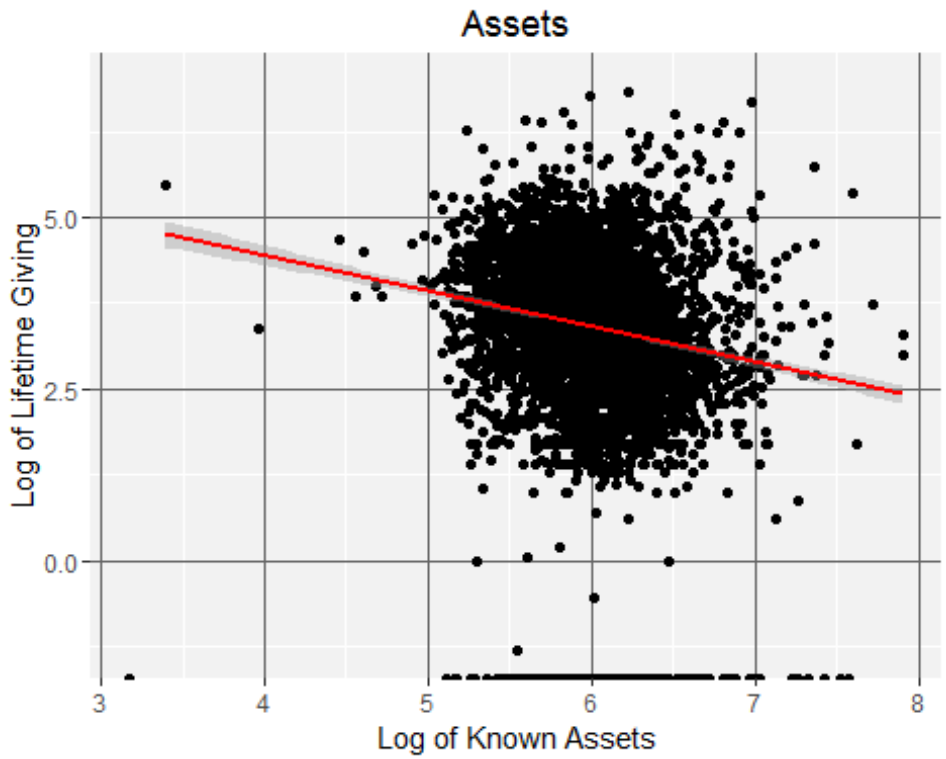
```
NoR.COR <- cor.test(TS$Number.of.relationships, TS$logGiving, method = "kenda
ll")
NoR.COR

##
##  Kendall's rank correlation tau
##
## data:  TS$Number.of.relationships and TS$logGiving
## z = 23.899, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2131592
```

The "Relationships" variable would be easier to understand in terms of donors if the kinds of relationships where broken out. Are parents of former students more generous, or are people whose parents were students? Are people without children more likely to become donors? A recommendation for future analyses is to separate this variable by type of relationship and look at each type independently from the others to see if any correlation exists.

Assets

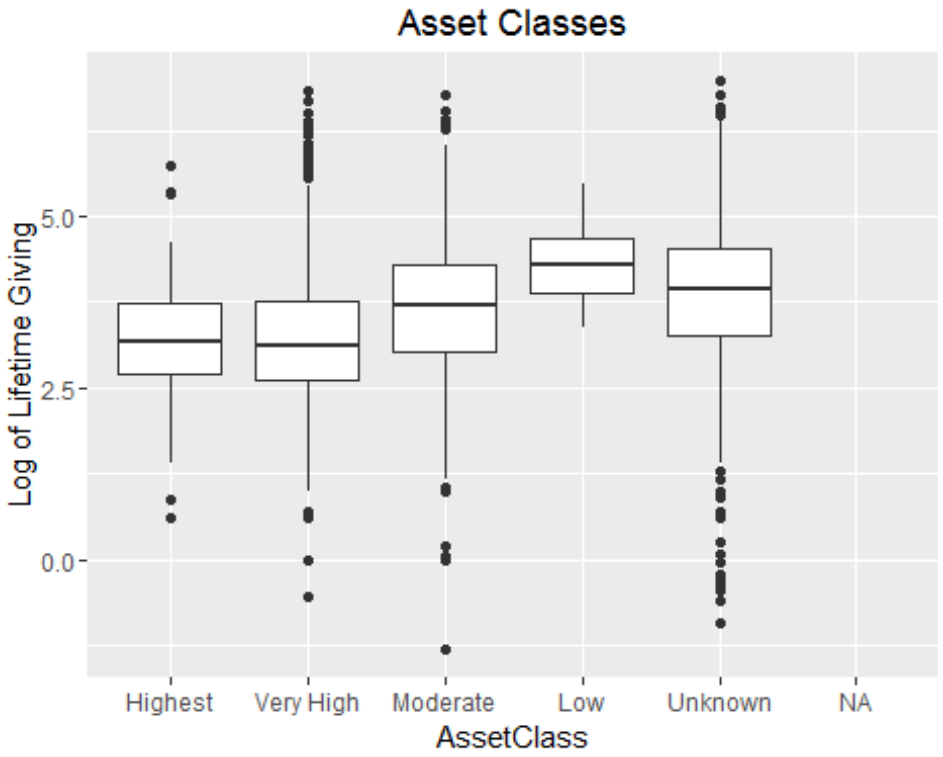
As with Lifetime Giving, the Asset variable was log transformed and the summary statistics are:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   3.176   5.763   6.026   6.018   6.250   7.903    2764
```



There are a couple of interesting things going on in this plot. The large mass in the center of the graph does not show any relationship between having assets and lifetime giving.

In order to try to understand assets as they relate to giving, the Assets variable was used to make five categories: Highest ($10M and above, 52 observations), Very High ($1M to $9,999,999, 2384 observations), Moderate ($100,000 to $999,999, 2103 observations), Low ($1 to $100,000, 11 observations) and Unknown, 2763 observations. Returning to the larger dataset (TS), the following box plot was made.

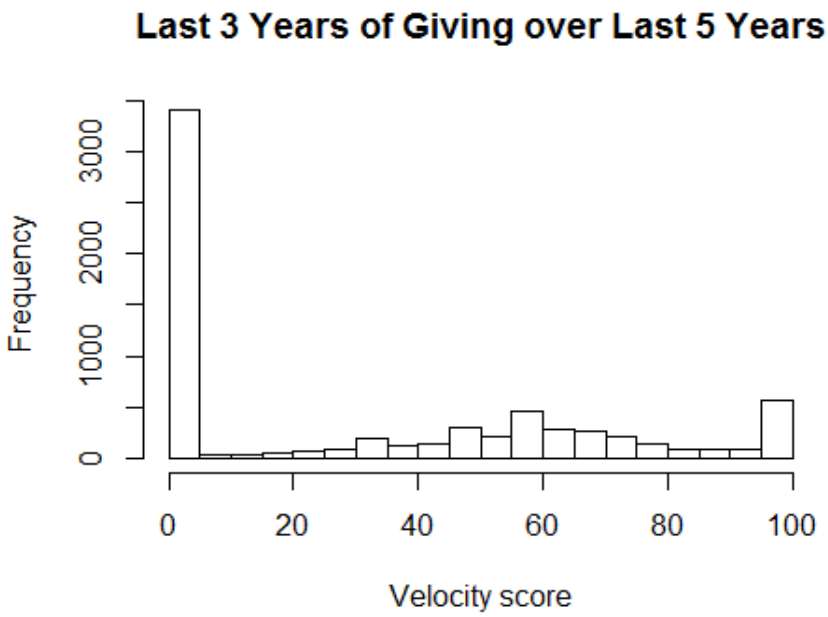


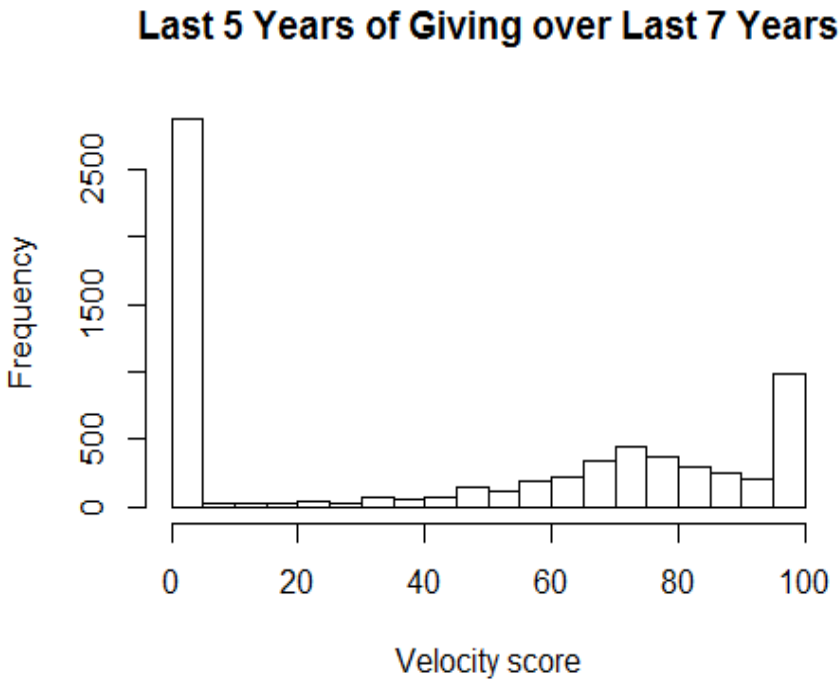
Of interest in this plot are the "Low" and "Unknown" boxes. Perhaps people with fewer assests are more focused in their giving. Or perhaps at least some of these people actually have assets that are not captured by this dataset (for example, observation 6222 has the highest giving amount in this dataset but no assets). It is likely that the highest donors also have the highest assets, but they might also have hidden those assets, so could be contained in the "Unknown" class in the plot. The variable AssetClass will be included in the regression model to see if it is significant in relation to giving.

Since the easiest asset data to find is the value of real estate, business and investment asset values could be a missing piece of information for some people in the "Low" and "Unknown" categories. Only more research on specific individuals could answer this question in future work.

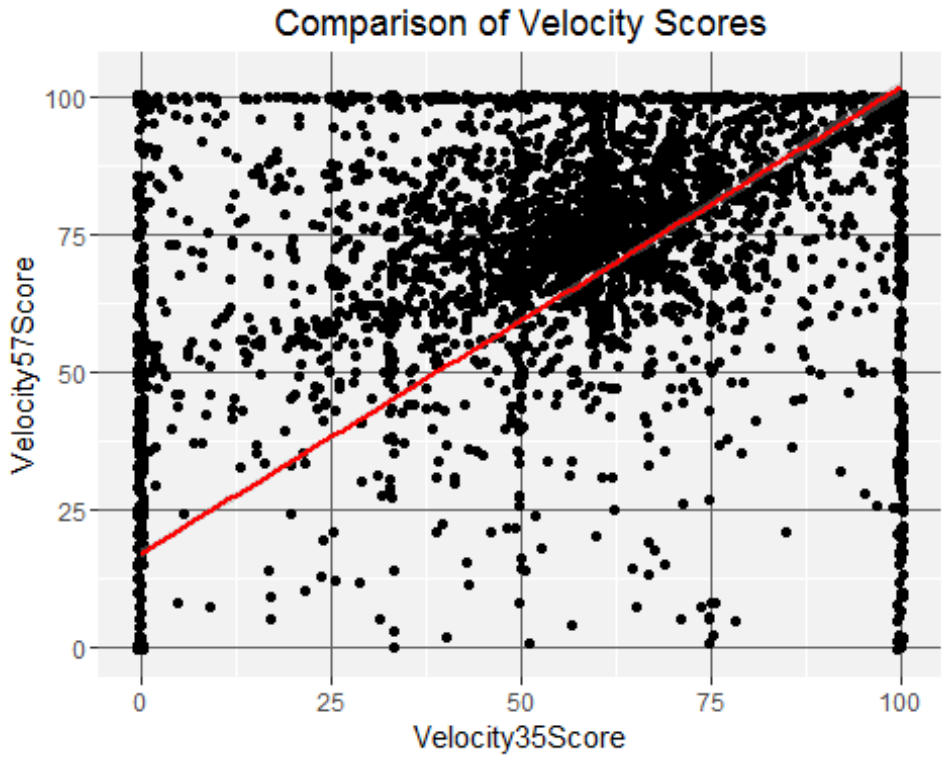
Velocity and Lifetime Giving

Velocity is a measure of the trajectory of recent giving. Literature on the subject uses two different methods of calculating velocity. The first method (the Velocity35Score) sums giving over the most recent three years and divides that number by the sum of giving for the past five years. The second method (Velocity57Score) sums the most recent five years and divides that by the sum of the most recent seven years. For the purposes of this study, it is calculated both ways to see which would be best to use in a regression model.





There is an interesting shift to the right from the first velocity plot to the second even though their shapes are close to the same. When compared against each other, the Velocity35Score has a positive relationship to the Velocity57Score.



The scatter plot is interesting in that it shows a positive relationship between the scores, but it does not say which one should be included in the regression model. Let's check wich one seems to have a higher correlation with lifetime giving using a Spearman's correlation since both varibles are skewed.

```
cor.test(TS$WSU.LIFETIME.GIVING, TS$Velocity35Score, method = "spearman")
```

```
## Warning in cor.test.default(TS$WSU.LIFETIME.GIVING, TS$Velocity35Score, :
## Cannot compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  TS$WSU.LIFETIME.GIVING and TS$Velocity35Score
## S = 2.4302e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5427463
```

```r
cor.test(TS$WSU.LIFETIME.GIVING, TS$Velocity57Score, method = "spearman")
```

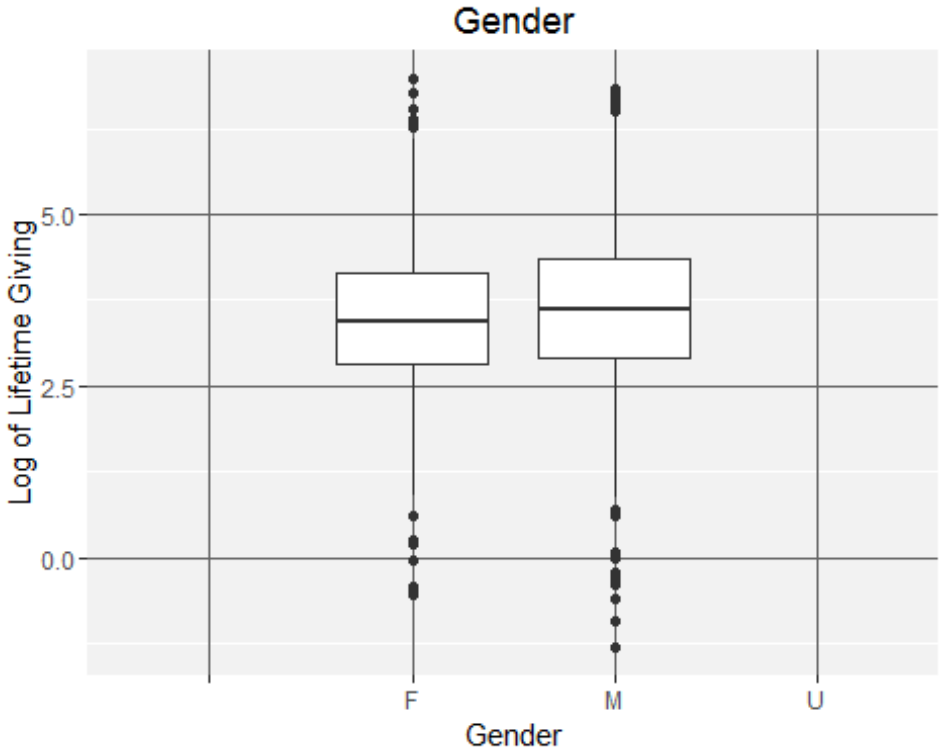
```
## Warning in cor.test.default(TS$WSU.LIFETIME.GIVING, TS$Velocity57Score, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  TS$WSU.LIFETIME.GIVING and TS$Velocity57Score
## S = 2.5554e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5192007
```

Both variables show a high Spearman rank correlation with lifetime giving. I will ue the Velocity35Score in the regression model since it's rank score is slightly higher and it has a less skewed distribution than the Velocity57Score.

Gender

While the mean giving for males is slightly higher than that of females, this box plot does not show a great difference in giving related to gender.



Since it is unlikely to have any predidictive value for giving, the variable GENDER will not be included in the regression model.

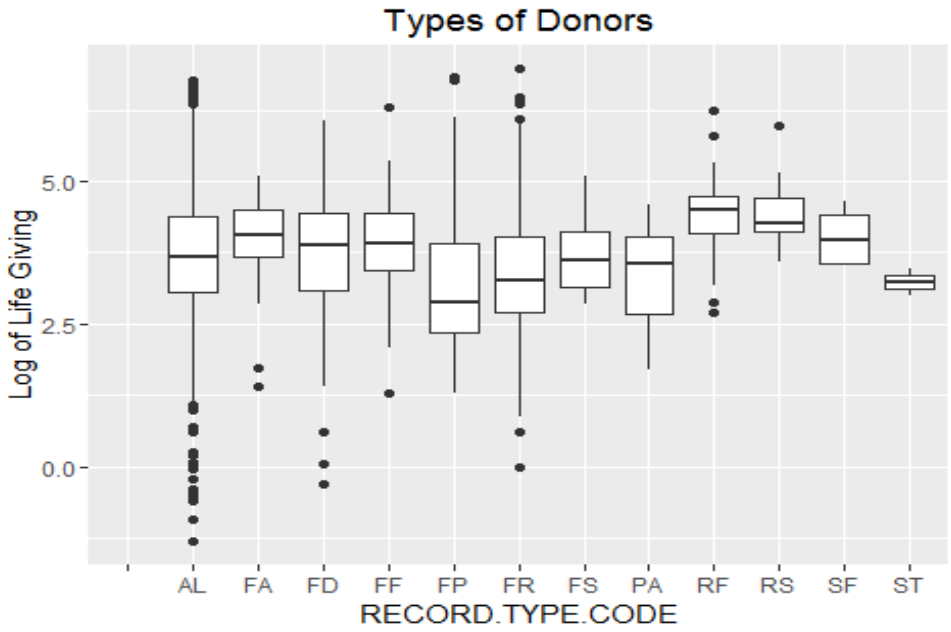
Types of donors in the sample

There are far more Alumni (AL) in the total sample than any other type of donor. The next largest group, Friend (FR), is someone who has a connection with the school through giving or other means, but who was never a student. Former Parents (FP) are the third largest group in this set. They are parents of a current or former student, but are not alumni themselves.

It seems straightforward to say that alumni make up our largest giving group, since they represent the bulk of the sample regardless of giving history. By themselves, alumni make up 63.8% of the TS sample. Combined with friends, they account for 87.1% of the sample.

```
relation.freq = table(TS$RECORD.TYPE.CODE)
relation.freq

##        AL   FA   FD   FF   FP   FR   FS   PA   RF   RS   SF   ST
##     1 4668   41  121   55  612 1706   14   27   52   13    4    3
```

Code	Description
AL	Alumnus
FA	Faculty
FD	Former Student (did not graduate with a degree)
FF	Former Faculty
FP	Former Parent
FR	Friend
FS	Former Staff
PA	Parent
RF	Retired Faculty
RS	Retired Staff
SF	Staff
ST	Student
---	---



In order to understand the influence of alumni status on lifetime giving, Welch's Two Sample t-test was performed on alumni and non-alumni giving.

```
Alum = TS$AlumniCode == 1
AlumGiving=TS[Alum,]$WSU.LIFETIME.GIVING  #Lifetime Giving by alumni

NoAlum = TS$AlumniCode == 0
AllElseGiving=TS[NoAlum,]$WSU.LIFETIME.GIVING  #Lifetime Giving by non-alumni

t.test(AlumGiving, AllElseGiving)

##
##  Welch Two Sample t-test
##
## data:  AlumGiving and AllElseGiving
## t = 1.1876, df = 4228, p-value = 0.2351
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4885.056 19897.300
## sample estimates:
## mean of x mean of y
##  40995.61  33489.49
```

Given the high p-value, the null hypotheses cannot rejected.

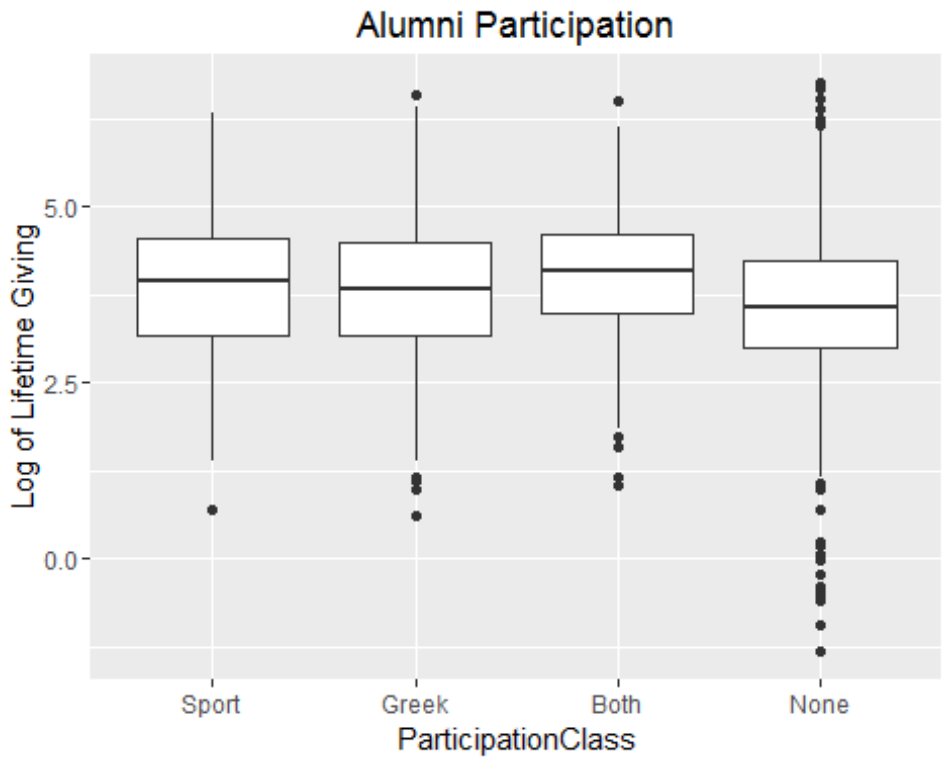
Alumni status might not appear to be influencing giving according to this test, but the AlumniCode variable will be included in the regression model even though other variables may turn out to be more significant.

This is one of the easiest datapoints to gather about a potential donor and might be useful in understanding why a potential donor might give when combined with other variables in future analyses.

Participation in a sport or a Greek chapter while a student

This variable only applies to alumni since it refers to activities while a student. A new dataset, TS6, restricts the sample to alumni only. This score is computed by assigning one "point" for membership in a Greek chapter or sports club, then totalling the points.

```
TS6 <-subset(TS, AlumniCode > 0)
```



There does not seem to be much difference between the medians of the categories based on this box plot. Welch's Two Sample t-test was performed on these variables to see if there was a meaningful difference between giving from those who participated in sports and those who joined a Greek chapter.

```
Spt = TS6$SportCode == 1
PlayedSport=TS6[Spt,]$WSU.LIFETIME.GIVING  #Lifetime Giving by those who play
ed Sports

NSpt = TS6$SportCode ==0
NoSport=TS6[NSpt,]$WSU.LIFETIME.GIVING #Lifetime Giving by those who didn't p
lay Sports

t.test(PlayedSport, NoSport)

##
##  Welch Two Sample t-test
##
## data:  PlayedSport and NoSport
## t = 0.72204, df = 382.62, p-value = 0.4707
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -15995.88  34562.29
## sample estimates:
## mean of x mean of y
##  52325.53  43042.33
```

```
Grk = TS6$GreekCode == 1
Greek=TS6[Grk,]$WSU.LIFETIME.GIVING  #Lifetime Giving by those who went Greek

NoGrk = TS6$GreekCode == 0
GDI=TS6[NoGrk,]$WSU.LIFETIME.GIVING  #Lifetime Giving by non-Greeks

t.test(Greek, GDI)
```

```
##
##  Welch Two Sample t-test
##
## data:  Greek and GDI
## t = 2.4534, df = 3630.4, p-value = 0.0142
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3273.922 29325.036
## sample estimates:
## mean of x mean of y
##  53878.48  37579.00
```

```
t.test(PlayedSport, Greek)
```

```
##
##  Welch Two Sample t-test
##
## data:  PlayedSport and Greek
## t = -0.1154, df = 456.72, p-value = 0.9082
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27998.61  24892.72
## sample estimates:
## mean of x mean of y
##  52325.53  53878.48
```

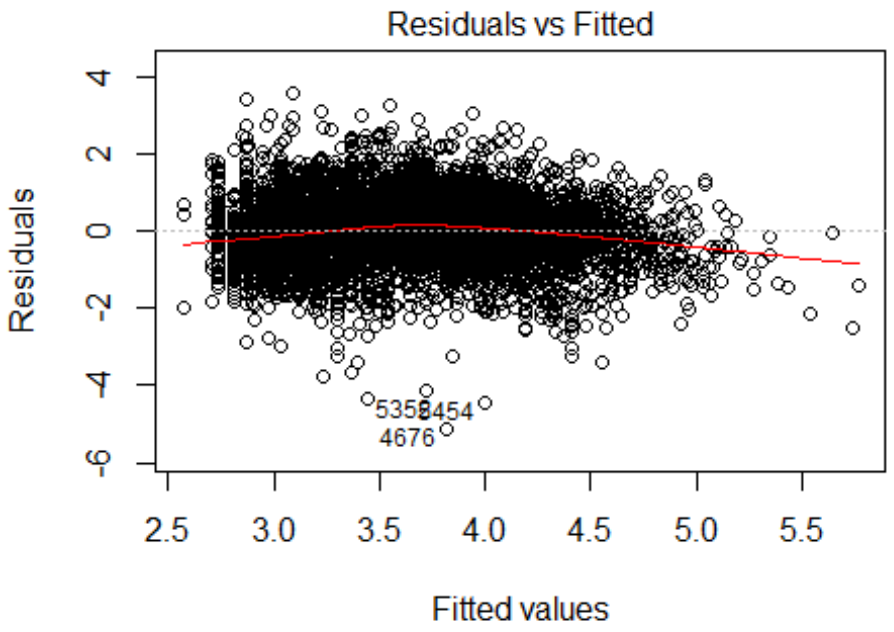
Given the high p-values, the null hypotheses cannot rejected. Both GreekCode and SportCode should not be expected to have a predictive influence on lifetime giving. But using both variables in the regression model is in agreement with literature discussing the effect on affinity on giving. People who participate in campus activities often seem to remain connected to the school and support it financially. The will be included, but are expected to show a lesser significance than other variables.

Regression Model

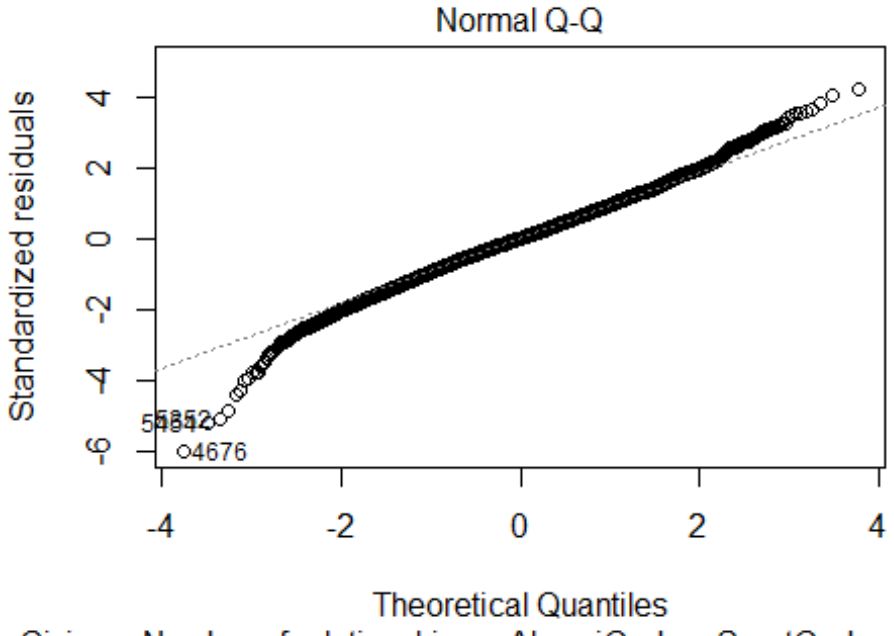
All of the data exploration has lead to deciding which variables are best suited for inclusion in a regression model.

These seven variables are included: logGiving (the dependent variable), Number.of.relationships, AlumniCode, SportCode, GreekCode, Velocity35Score, and AssetClass, in a regression model.

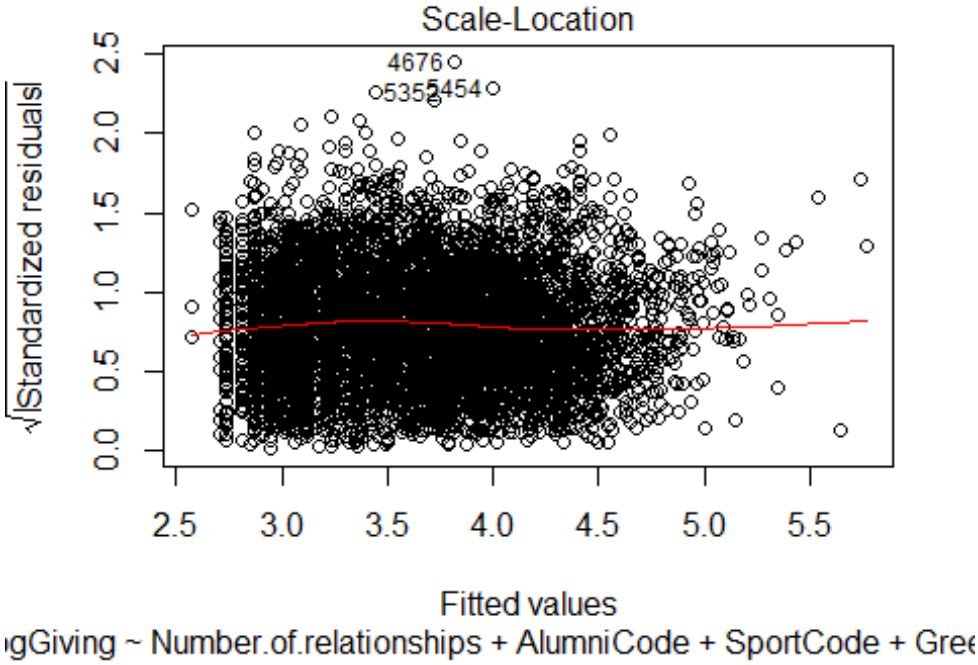
```
TS$AssetClass2 <- as.numeric(TS$AssetClass)
TS$logGiving[which(!is.finite(TS$logGiving))] = NA
RegMod <-lm(logGiving~Number.of.relationships+AlumniCode+SportCode+GreekCode+
Velocity35Score+AssetClass2, data=TS)

summary(RegMod)

##
## Call:
## lm(formula = logGiving ~ Number.of.relationships + AlumniCode +
##     SportCode + GreekCode + Velocity35Score + AssetClass2, data = TS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1151 -0.5160  0.0006  0.5362  3.5781
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.404509   0.034372  69.955  < 2e-16 ***
## Number.of.relationships 0.140236   0.007086  19.791  < 2e-16 ***
## AlumniCode               0.077489   0.025583   3.029 0.002465 **
## SportCode                0.131743   0.049729   2.649 0.008088 **
## GreekCode                0.100598   0.026849   3.747 0.000181 ***
## Velocity35Score          0.009697   0.000313  30.984  < 2e-16 ***
## AssetClass2              0.163537   0.008652  18.901  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8505 on 6171 degrees of freedom
##   (1139 observations deleted due to missingness)
## Multiple R-squared:   0.29,  Adjusted R-squared:  0.2893
## F-statistic: 420.1 on 6 and 6171 DF,  p-value: < 2.2e-16

plot(RegMod)
```

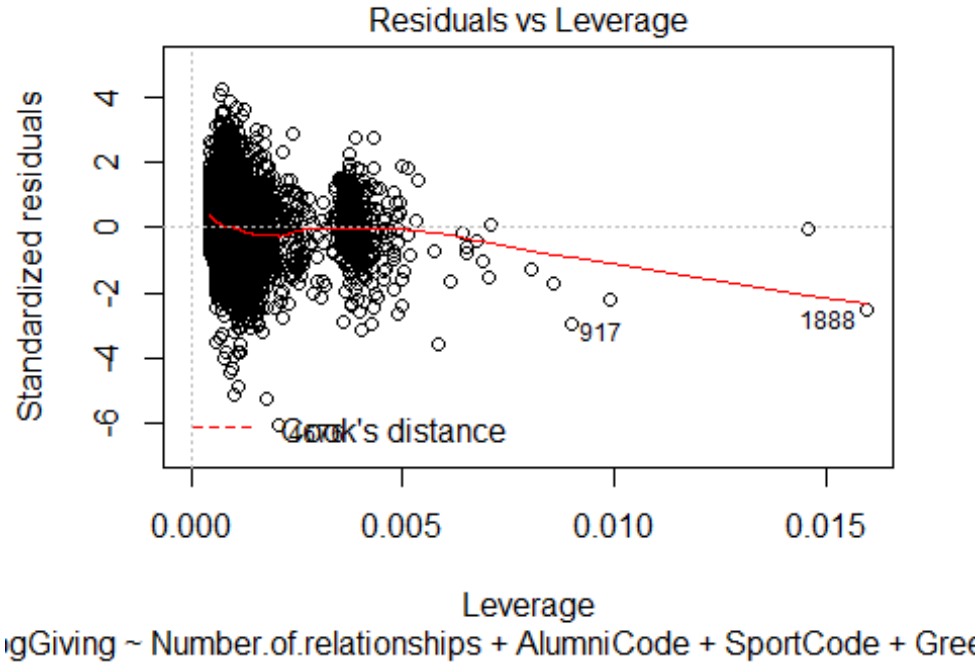


gGiving ~ Number.of.relationships + AlumniCode + SportCode + Gree



gGiving ~ Number.of.relationships + AlumniCode + SportCode + Gree





These plots seem to show that the logistic regression model accounts for most of the observations, but there are data points far from the lines, especially the three numbered points.

```
Outlier<- TS[c(3625,2822,6222),]
Outlier
```

```
##      IDCode WSU.LIFETIME.GIVING WSU.YEARS.OF.GIVING  ASSETS
## 3625   1014             5889463                  22  986852
## 2822     66             6657303                  26 1656000
## 6222    988             9672888                  35      NA
##      RECORD.TYPE.CODE Number.of.relationships Gender Velocity35Score
## 3625               AL                       4      M              64
## 2822               FP                       3      M              41
## 6222               FR                       4      F              16
##      Velocity57Score AlumniCode SportCode GreekCode ParticipationScore
## 3625              86          1         0         0                   0
## 2822              81          0         0         0                   0
## 6222              55          0         0         0                   0
##      AssetClass ParticipationClass logGiving logAssets AssetClass2
## 3625   Moderate               None  6.770076  5.994252           3
## 2822  Very High               None  6.823298  6.219060           2
## 6222    Unknown               None  6.985556        NA           5
```

These three observations were outliers because they contained the three largest lifetime giving amounts. Only one of these is an alum, and the person who has given the most has no assets in this dataset.

Looking at the residual plots above, it seems like the variables AlumniCode, WSU.YEARS.OF.GIVING, and Number.of.relationships might predict the giving of most of the donors, they do not characterize the larger givers. Future research should look at other variables available in the donor database to see if they can be used to refine the model and predict larger donations.

Recommendations for Future Analyses

First Recommendtion

Find a way to determine when an individual made their first gift and see if that predicts lifetime giving. Obviously, the longer someone has been giving, the higher their lifetime giving might be. But does age at first gift predict larger gifts? There are people in the larger donor database who have been giving for 50 years or more without becoming major donors, so years of giving does not tell the whole story.

Second Recommendation

Break the Number.of.relationship variable into separate variables for grandparents, spouses, children, and so on, to see if any type of family relation has any influence on lifetime giving.

Third Recommendation

It would be interesting to examine how many years ago donors began giving. It would also be interesting to know their employer and major, if they are alumni. Are they Boeing executives who began giving once they became executives? Are these Microsoft employees who have begun donating as soon as they began their working careers?Is there a major or field of study more likely to result in a major gift? Does long term giving indicate a likelihood of including the school in their will? These are questions for future research.

Fourth Recommendation

The variables chosen for this project were relatively easy to collect from our donor database. Future analysis should explore other variables that might have more predictive ability but may be more difficult to collect and use. For instance, a Recency Score (how many years ago was the largest gift made?) and Largest Gift Score might be significant predictors.

Fifth Recommendation

It would be very interesting to look at covariance and if any of these variable are working together to influence giving. There might be variables that turn out to be proxies for data points that are not represented in the donor database. Does living on Bainbridge Island stand as a proxy for income or investment assets?

Project Conclusions

There are many other kinds of statitical tests that could be performed on this dataset, but the results would likely be the same. There are no variables that have a strong enough predictive relationship to lifetime giving to build a useful predictive model. The regression models show alumni status, years of giving, and number of relationships have a predictive significance on lifetime giving but the range of predicted values is very large. The addition of other variables, and the removal of non-significant ones, might improve the model.

The graphs of most of the variables in the TS dataset pointed to some correlation, or lack thereof, with lifetime giving. They helped sort out variables like Gender that would not

contribute to the model. The first Asset scatter plot also shows that this dataset is very noisy. Removing those obervations that had no giving would clear up most of the noise.

This study has been only a starting point in examining and understanding data that can be extracted from the donor database and exploring how it can be used to create a predictive model.