

# DataScience Capstone Project

Missy Lee  
11/15/2016

## 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?

State funding has declined over the past decade at the same time as the need for new facilities and programs has increased. Private philanthropy is looked upon to help fill the funding gap. We have very loyal and generous alumni who made the last campaign a success by donating very large amounts. These people may not have the inclination or financial ability to make similar gifts in the next campaign. Results from this and future analytical efforts will be used to identify people who may be likely major gift donors, but who are presently not identified as major gift prospects.

The dataset for this study was randomly sampled for 3036 individuals 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.

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.

```
TS <- read.csv("~/GitHub/Mlee-Data-Science-Capstone-Project/TS.csv", header=TRUE)
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
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
6 6 6 6 6 6 6 6 ...
## $ ParticipationClass       : Factor w/ 5 levels "", "Both", "Greek",...: 2 2 2
2 2 2 2 2 2 2 ...
```

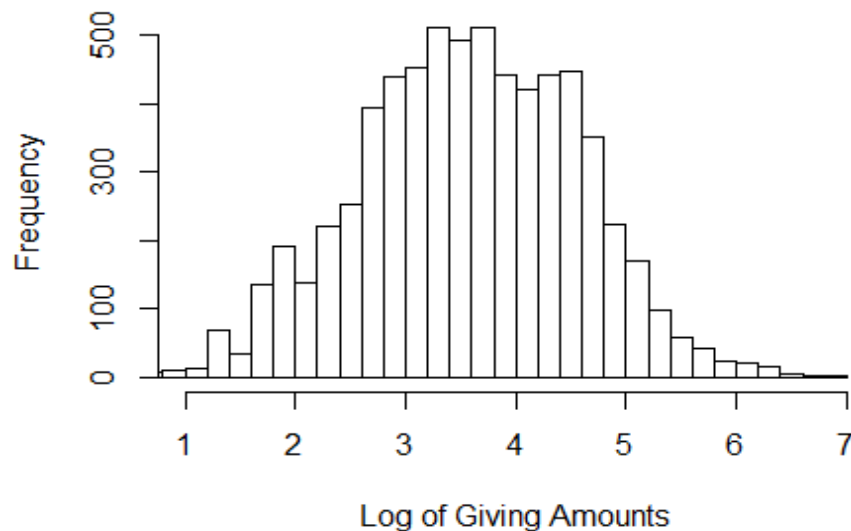
## Lifetime Giving

In order to show a normal distribution, the Lifetime Giving data needed to be log10 transformed.

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

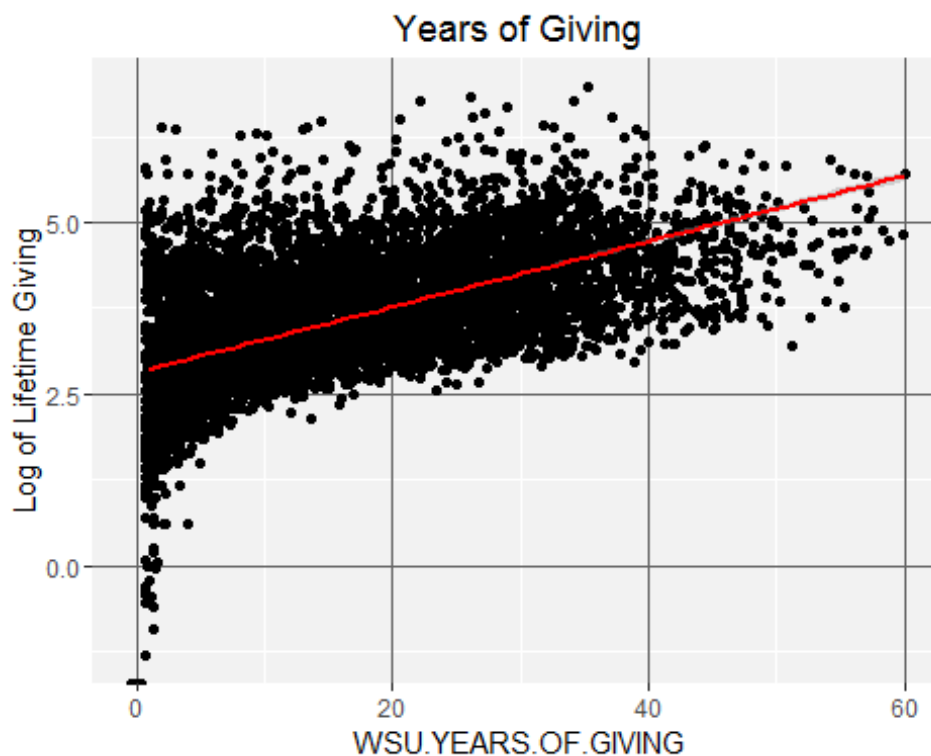
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##      -Inf   2.653   3.447   -Inf   4.235   6.986         1
```

## Distribution of Lifetime Giving Amounts



Since the goal of future modeling will be to predict the likelihood of a person donating money, Lifetime Giving is assigned the role of dependent variable in this study. The histogram of the log of lifetime giving amounts shows a normal distribution so it is useable as a variable to contrast with other variables.

## Years of Giving

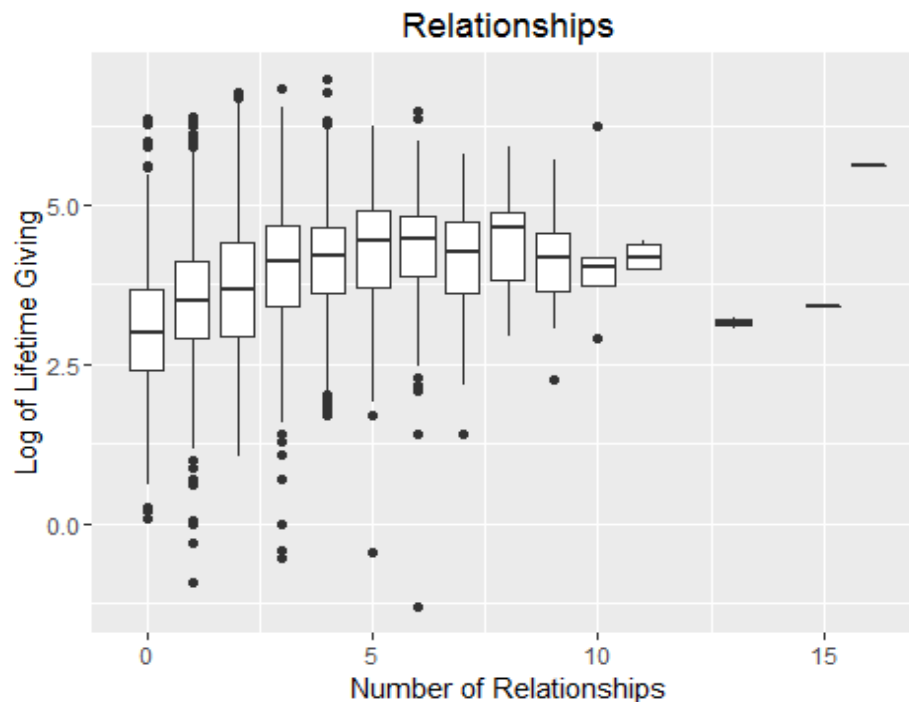


The Years of Giving scatter plot shows some large donors have been giving for between 20 and 40 years. There are also donors who have given longer, but less overall. It is to be expected that giving over long amounts of time does add up to a large amount over a lifetime, and this plot does show that positive relationship.

The variable WSU.YEARS.OF.GIVING will be 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



This plot seems to indicate a correlation between lifetime giving and a low to moderate number of other family members who attended the university. Relationships in the donor database include parents, grandparents, aunts, uncles, siblings, as well as children.

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.

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

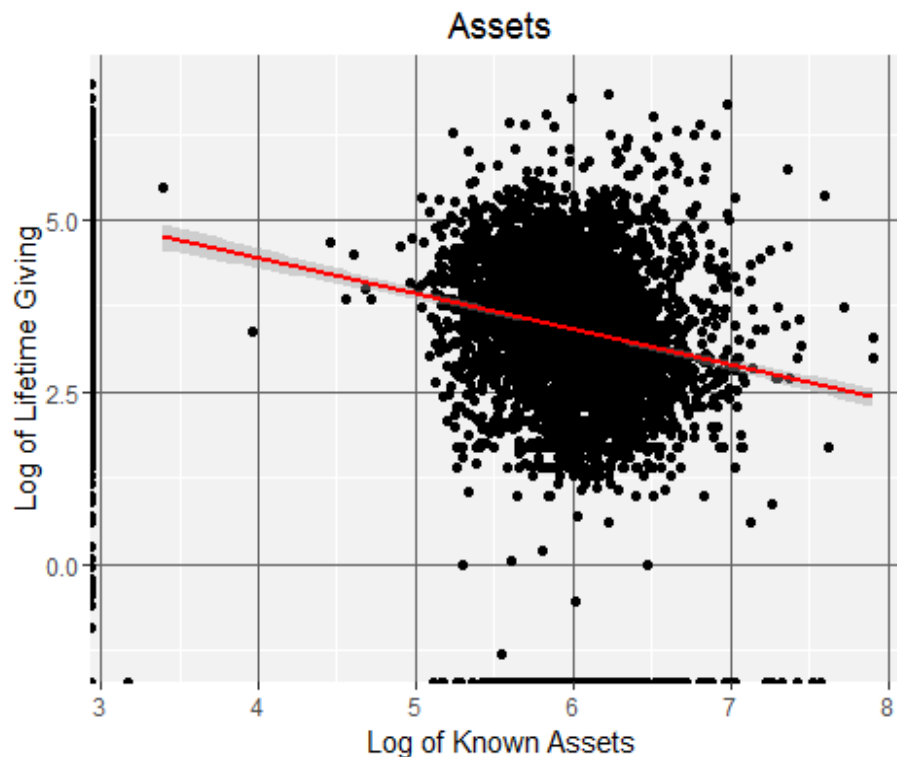
## Kendall's rank correlation tau
##
## data: TS$Number.of.relationships and TS$logGiving
## z = 23.899, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.2131592
```

The "Relationships" variable would be easier to understand in terms of donors if the kinds of relationships were 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:

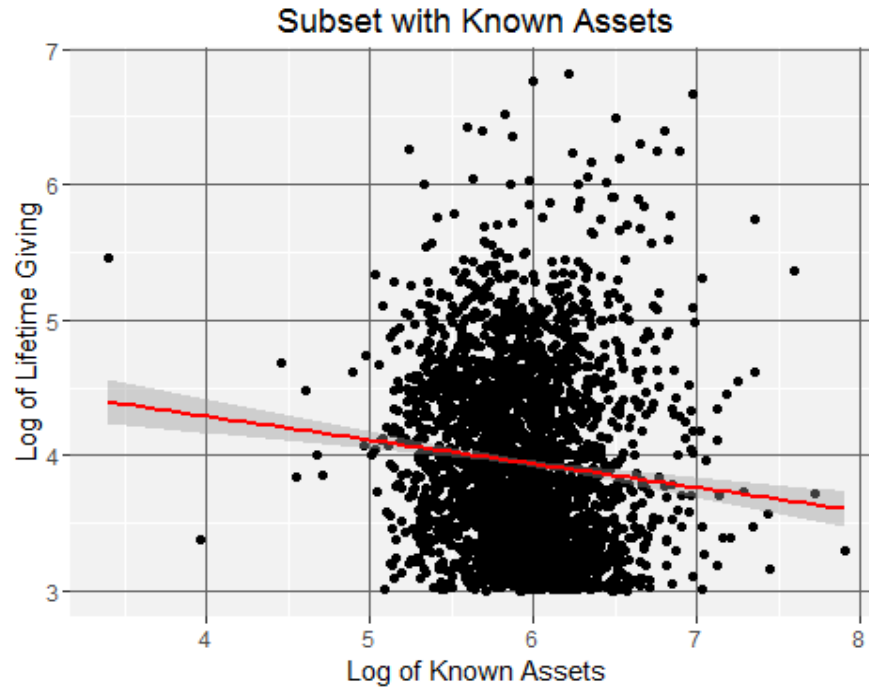
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	-Inf	-Inf	5.694	-Inf	6.112	7.903	1



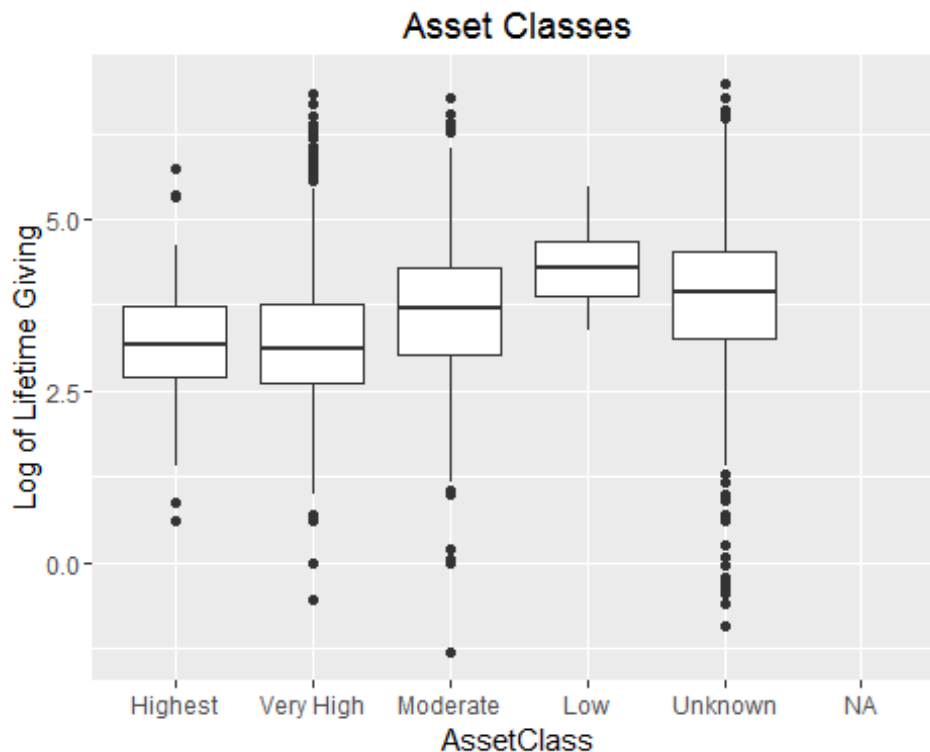
There are a couple of interesting things going on in this plot. The large mass in the center of the graph does not show any relationship between having assets and lifetime giving. The solid line on the left of giving by those with no known assets and no giving could be a data collection issue, since asset data has not been collected for people who have given nothing or small amounts. The data will be subsetting to exclude them and plotted again.

Removing the records without assets data and with less than \$1,000 giving results still shows a negative relationship, and there are quite a few points well away from the regression line that correspond to high giving amounts.

```
TS4 <- subset(TS, ASSETS > 0 & WSU.LIFETIME.GIVING > 1000)
subassets <- ggplot(TS4, aes(x=logAssets,y=logGiving))+geom_jitter()+stat_smooth(
  method="lm", col="red")+ theme(panel.background = element_rect(fill = "grey95"),
  panel.grid.major = element_line(colour = "grey40"))
subassets <- subassets + scale_x_continuous(name="Log of Known Assets") + scale_y_continuous(
  name = "Log of Lifetime Giving")
subassets <- subassets + ggtitle("Subset with Known Assets")
subassets
```



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 assets 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. For this reason, the variable ASSETS will be included in the regression model.

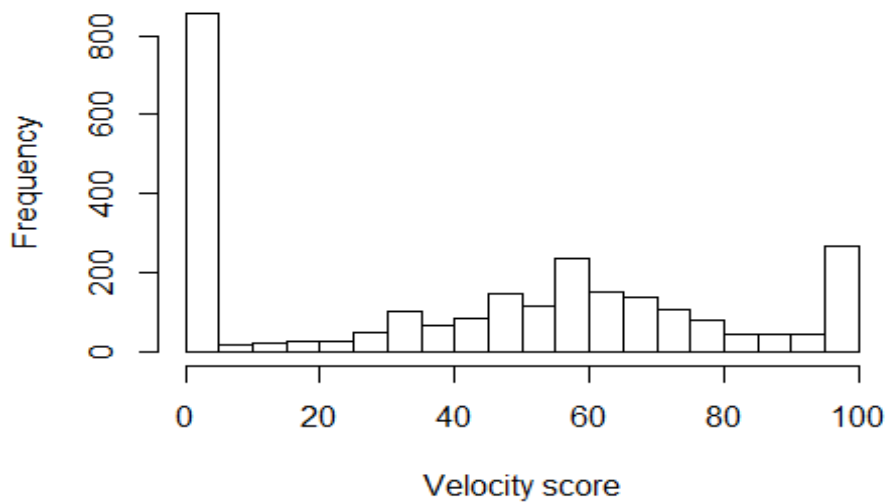
Since the easiest asset data to find is the value of real estate, business and investment asset values could be a missing piece of information for some people in the "Low" and "Unknown" categories. Only more research on specific individuals could answer this question.

But this type of research is generally only done for people who are already major donors. Real estate values are public information obtainable through county assessor records and websites like Zillow, but people often put their real estate into a family or private trust, making it harder to find. Business and investment data are harder to find and generally not public information. Stock holdings and executive compensation are likewise difficult to value.

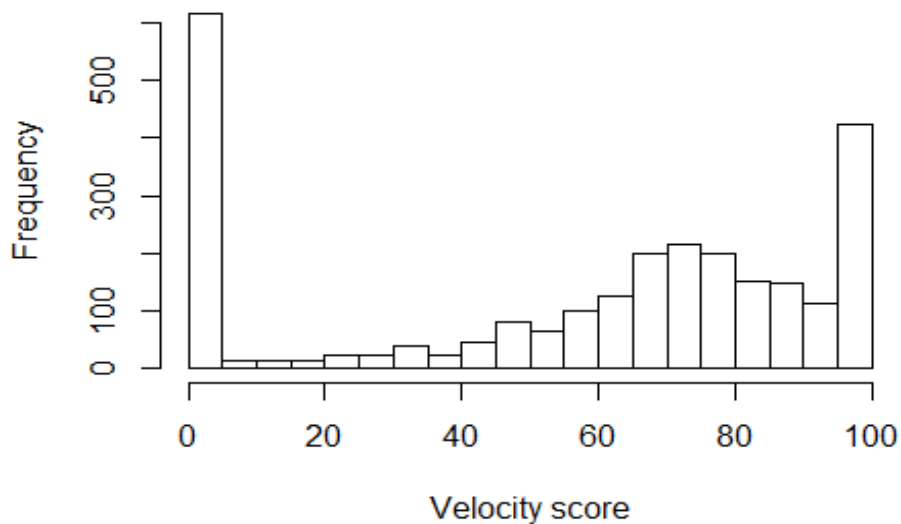
## Velocity and Lifetime Giving

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.

### Last 3 Years of Giving over Last 5 Years

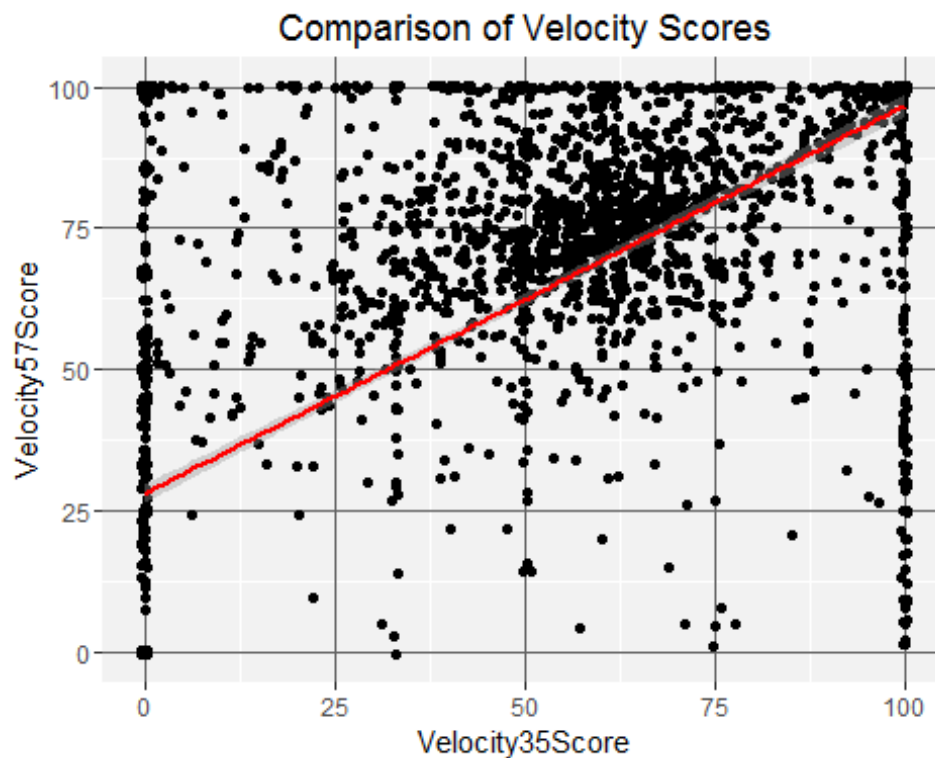


### Last 5 Years of Giving over Last 7 Years





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. Correlation tests might help.

```
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

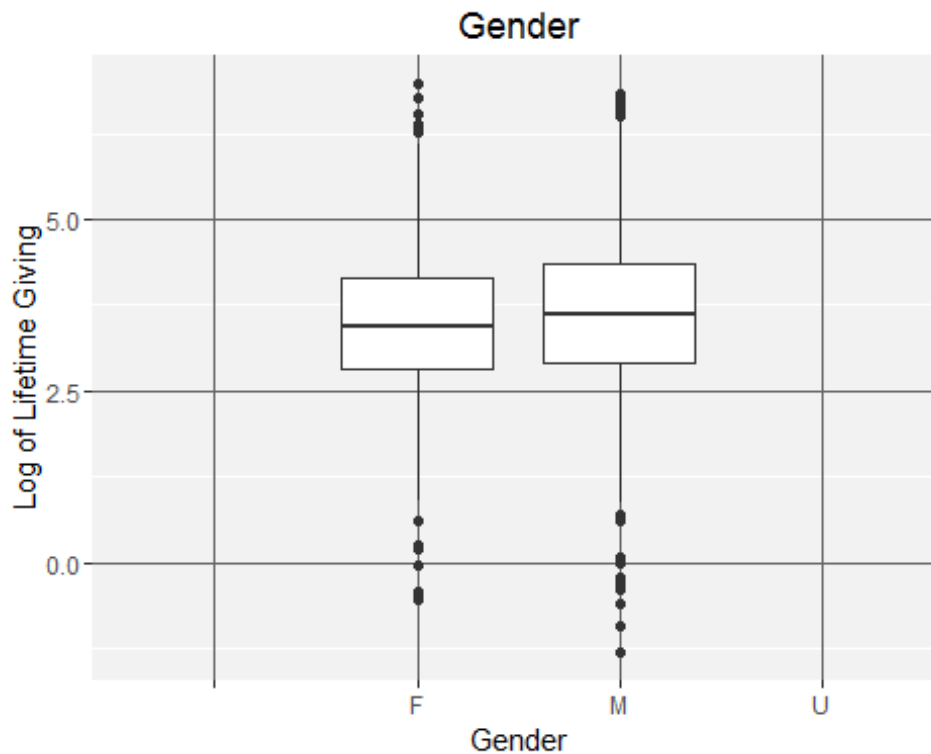
cor.test(TS$WSU.LIFETIME.GIVING, TS$Velocity57Score, method = "spearman")

## 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
```

The rho value for each velocity score, contrasted against lifetime giving, are quite high. The Velocity35Score will be included in the regression model since its r value is slightly better (0.543 versus 0.519 for Velocity57Score).

## Gender

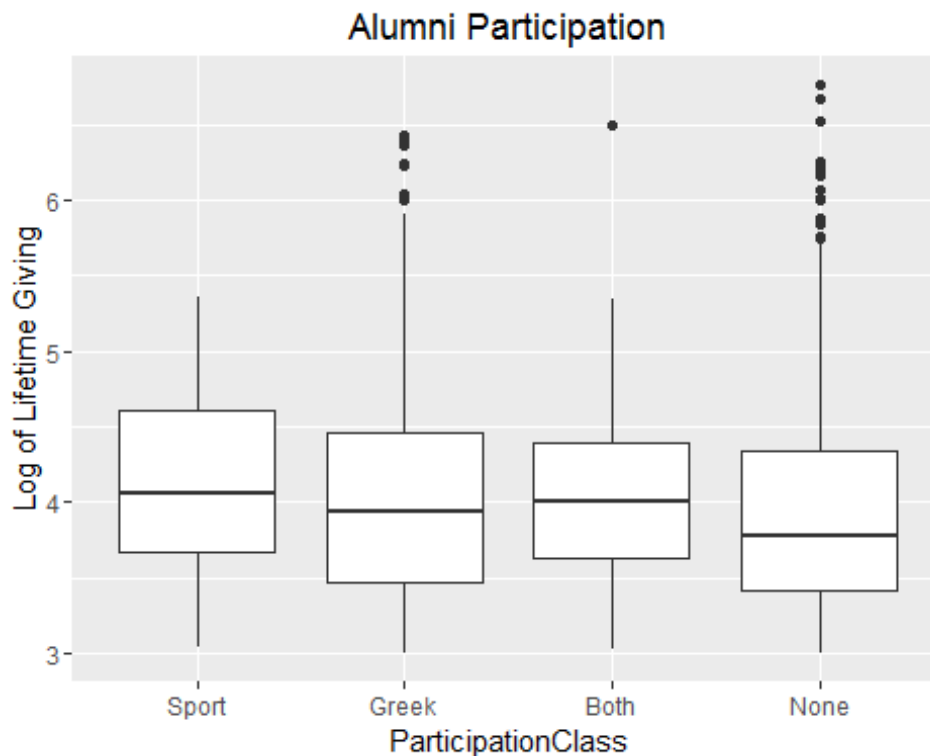


This box plot does not show a great difference in giving related to gender. Since it is unlikely to have any predictive value for giving, the variable GENDER will not be included in the regression model.

## 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 to alumni only, who have assets and have given more than \$1000. This score is computed by assigning one "point" for membership in a Greek chapter or sports club, then totalling the points.

```
TS6 <-subset(TS, ASSETS > 0 & WSU.LIFETIME.GIVING > 1000 & 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 = -0.13391, df = 182.03, p-value = 0.8936
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -43977.24 38387.18
## sample estimates:
## mean of x mean of y
## 47142.42 49937.45

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 = 0.2546, df = 1921.5, p-value = 0.7991
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -18794.32 24402.10
## sample estimates:
## mean of x mean of y
## 51507.11 48703.22

t.test(PlayedSport, Greek)

##
## Welch Two Sample t-test
##
## data: PlayedSport and Greek
## t = -0.20281, df = 204.98, p-value = 0.8395
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -46796.77 38067.38
## sample estimates:
## mean of x mean of y
## 47142.42 51507.11

```

Given the differences in the means of these samples and the high p-values, both GreekCode and SportCode should be part of the regression model. This result 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.

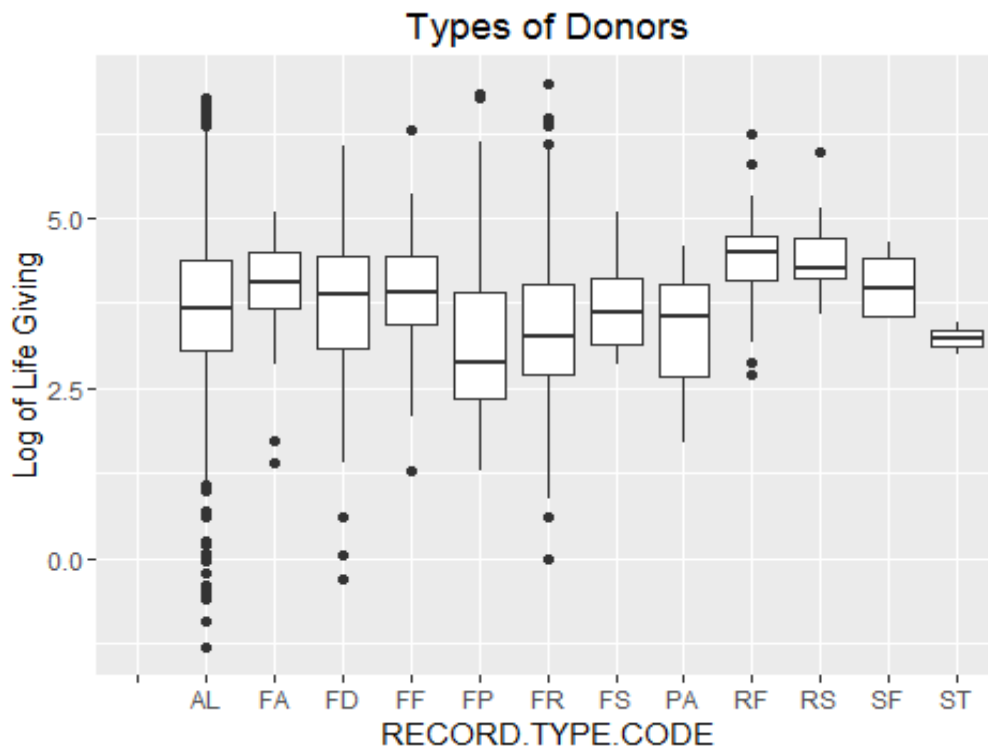
## Types of donors in the sample

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.

```
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
```

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. "Friend" is defined as someone who has donated or done business with the university but who did not go to school here.

The AlumniCode variable will be included in the regression model. It 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.

## Logistic Regression Model

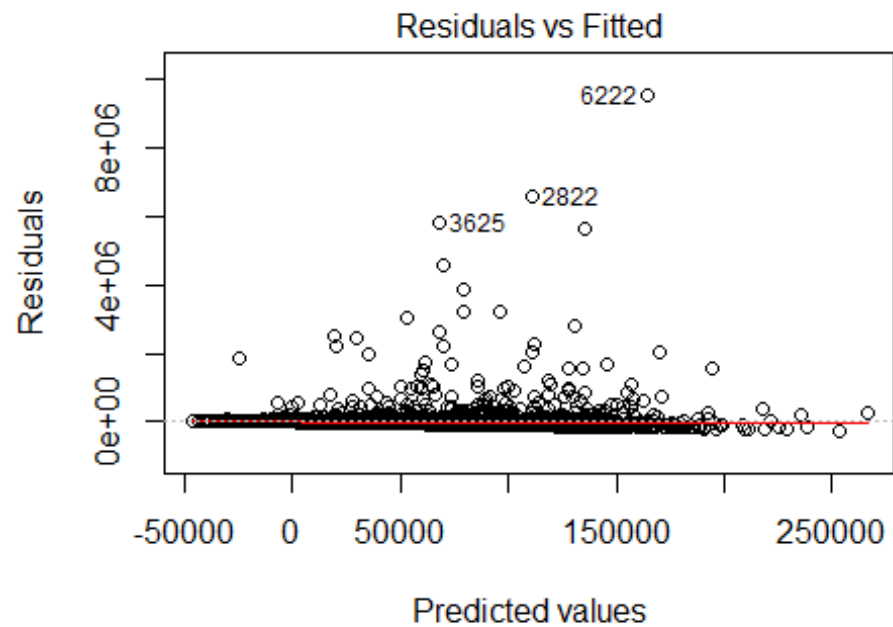
Logistic regression was chosen because monetary values like WSU.LIFETIME.GIVING need to be handled as a log in order to contrast it with other variables.

When these seven variables are included: WSU.LIFETIME.GIVING (the independent variable), WSU.YEARS.OF.GIVING, AssetClass, Number.of.relationships, AlumniCode, SportCode, GreekCode, Velocity35Score, in a logistic regression model, then AlumniCode along with WSU.YEARS.OF.GIVING, and Number.of.relationships are significant in relation to Lifetime Giving. The p-value is less than 0.05 means the null hypothesis can be rejected.

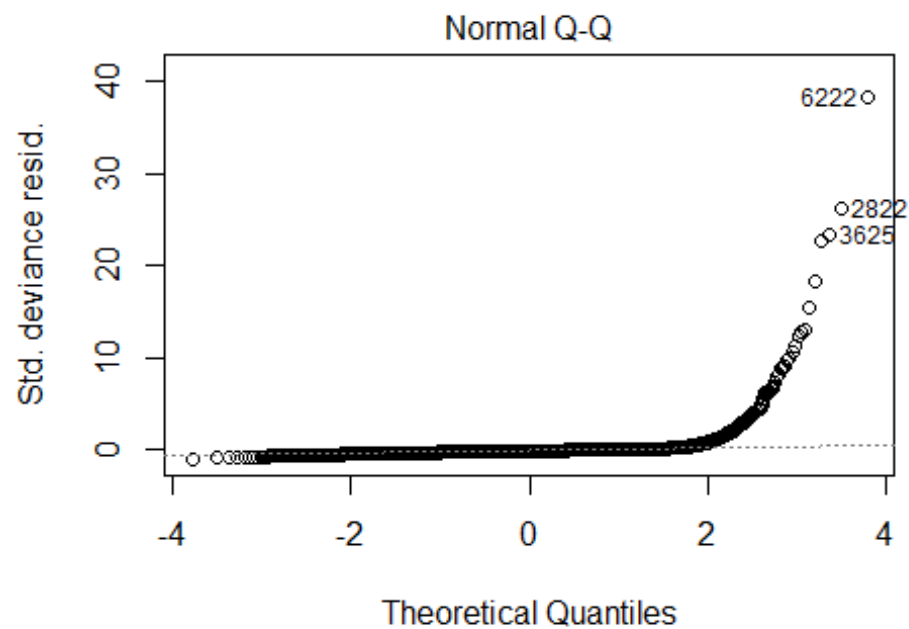
```
AssetClass2 <- as.numeric(TS$AssetClass)
RegMod <- glm(WSU.LIFETIME.GIVING~WSU.YEARS.OF.GIVING+AssetClass+Number.of.relationships+AlumniCode+SportCode+GreekCode+Velocity35Score, data=TS)
summary(RegMod)

## Call:
## glm(formula = WSU.LIFETIME.GIVING ~ WSU.YEARS.OF.GIVING + AssetClass +
##      Number.of.relationships + AlumniCode + SportCode + GreekCode +
##      Velocity35Score, data = TS)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -247657  -47107  -21750   1211  9508889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      852.04   36794.27   0.023   0.982
## WSU.YEARS.OF.GIVING    2812.28    305.61   9.202 < 2e-16 ***
## AssetClassVery High   -3958.85   37127.18  -0.107   0.915
## AssetClassModerate   -15236.11   37295.58  -0.409   0.683
## AssetClassLow        -12565.61   90829.78  -0.138   0.890
## AssetClassUnknown     10728.55   37218.68   0.288   0.773
## Number.of.relationships 13364.71    2071.38   6.452 1.18e-10 ***
## AlumniCode          -34870.28    7917.86  -4.404 1.08e-05 ***
## SportCode           2223.03    14525.71   0.153   0.878
## GreekCode            414.43     7830.04   0.053   0.958
## Velocity35Score       33.08       96.50   0.343   0.732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 61893755928)
##
##      Null deviance: 4.1308e+14  on 6479  degrees of freedom
## Residual deviance: 4.0039e+14  on 6469  degrees of freedom
## (837 observations deleted due to missingness)
## AIC: 179422
##
## Number of Fisher Scoring iterations: 2
```

```
plot(RegMod)
```

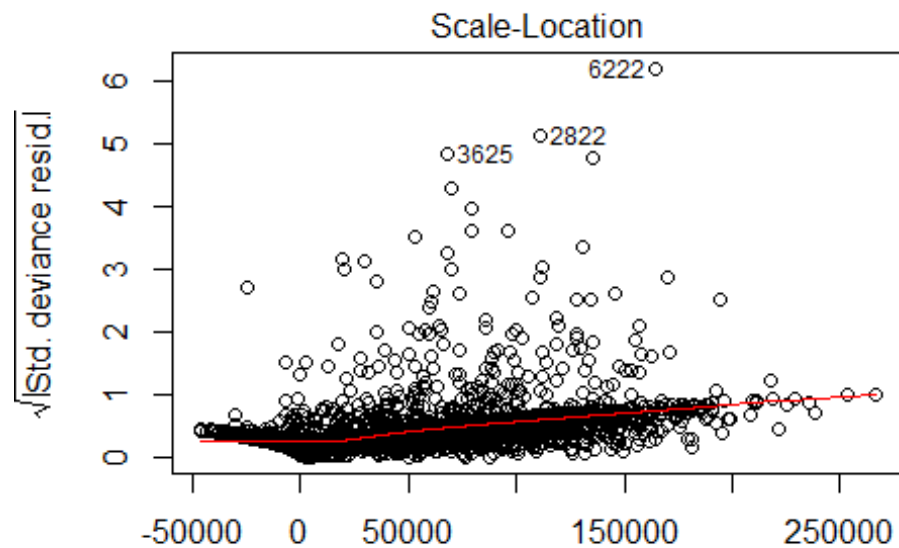


```
US.LIFETIME.GIVING ~ WSU.YEARS.OF.GIVING + AssetClass + Nurr
```



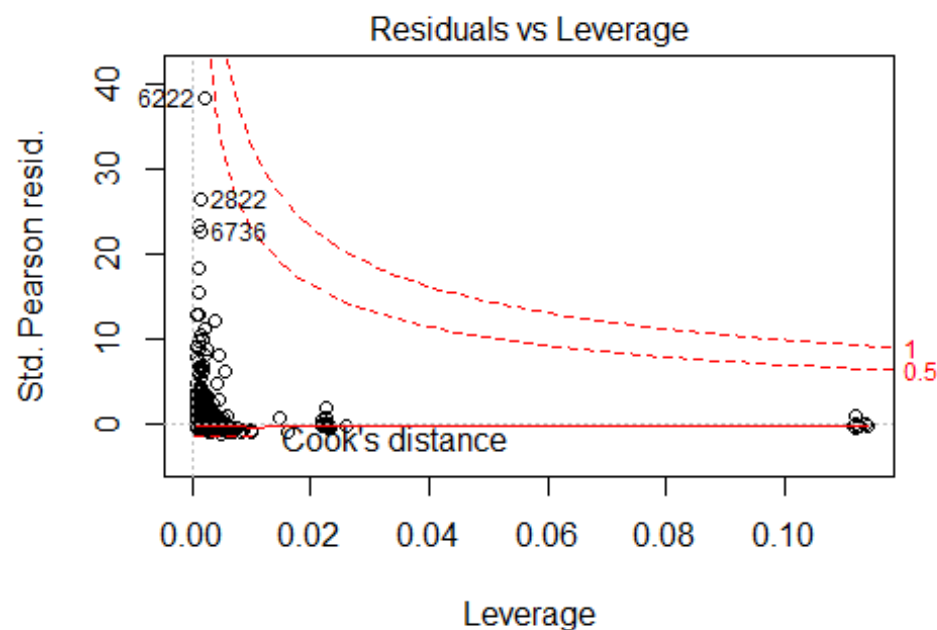
```
US.LIFETIME.GIVING ~ WSU.YEARS.OF.GIVING + AssetClass + Nurr
```





Predicted values

$\text{SU.LIFETIME.GIVING} \sim \text{WSU.YEARS.OF.GIVING} + \text{AssetClass} + \text{Nurr}$



Leverage

$\text{SU.LIFETIME.GIVING} \sim \text{WSU.YEARS.OF.GIVING} + \text{AssetClass} + \text{Nurr}$

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	0

##	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
## 3625	Moderate	None	6.770076	5.994252
## 2822	Very High	None	6.823298	6.219060
## 6222	Unknown	None	6.985556	-Inf

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 and is a friend, not an alum.

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 Recommendation

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 statistical 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 observations 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.