DataScience Capstone Project

Missy Lee

11/8/2016

Introduction

The dataset for this study was 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=T
RUE)
str (TS)
```

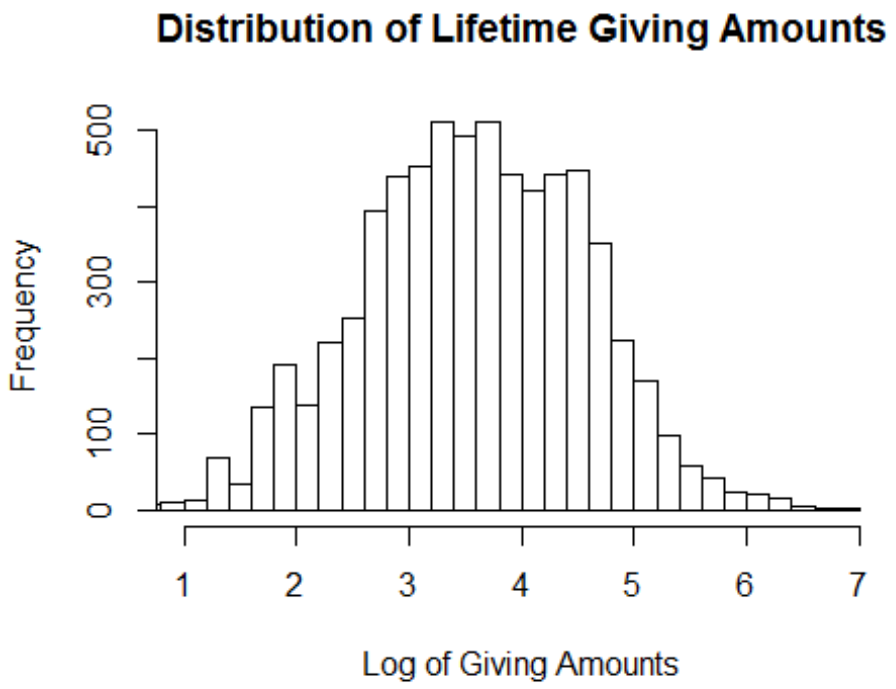
```
## 'data.frame':    7317 obs. of  15 variables:
##  $ IDCode                : int  404 3198 3592 3085 7059 112 4832 3245 361
9 6147 ...
##  $ WSU.LIFETIME.GIVING   : num  17712 19475 6080 24155 0 ...
##  $ WSU.YEARS.OF.GIVING   : int  29 20 25 22 0 21 28 18 31 2 ...
##  $ ASSETS                : 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
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.

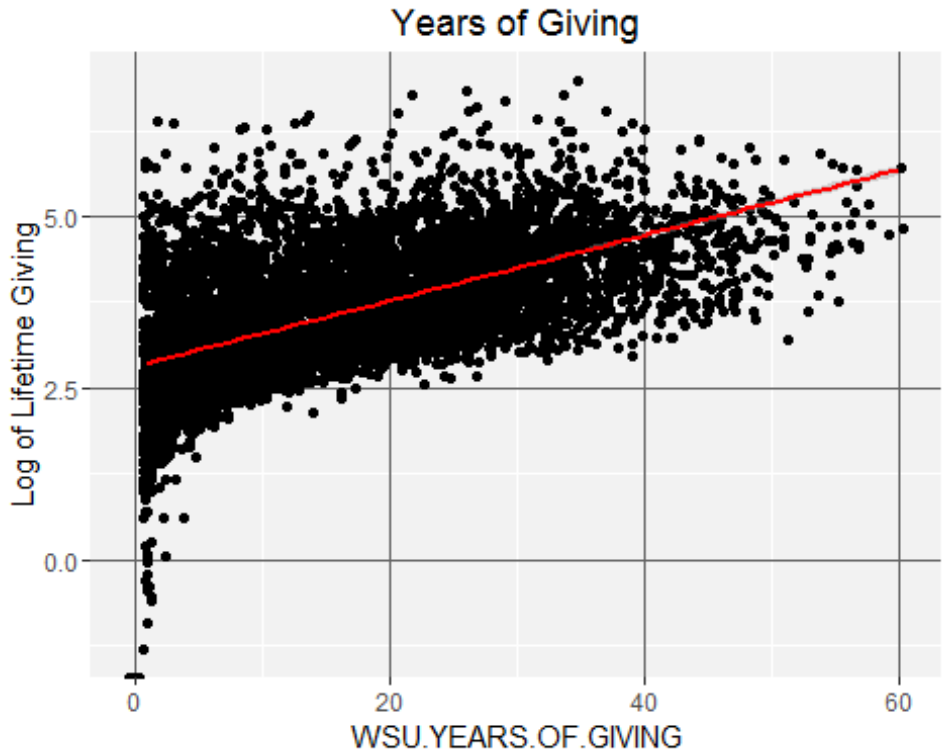
```
 TS$logGiving <- log10(TS$WSU.LIFETIME.GIVING)
summary(TS$logGiving)

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



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, but it is shifted to the right.

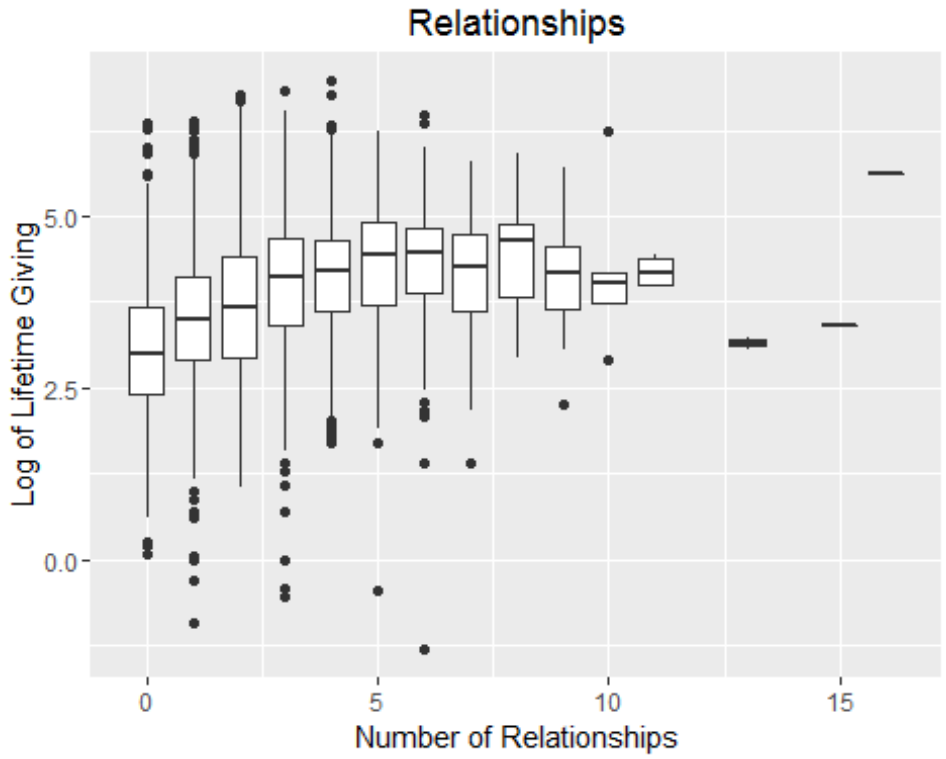
Years of Giving



The "Years of Giving" scatter plot shows some large donors have been giving for between 20 and 40 years. The variable WSU.YEARS.OF.GIVING will be included in the regression model since it seems to influence Lifetime Giving.

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? These are questions for future research.

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.

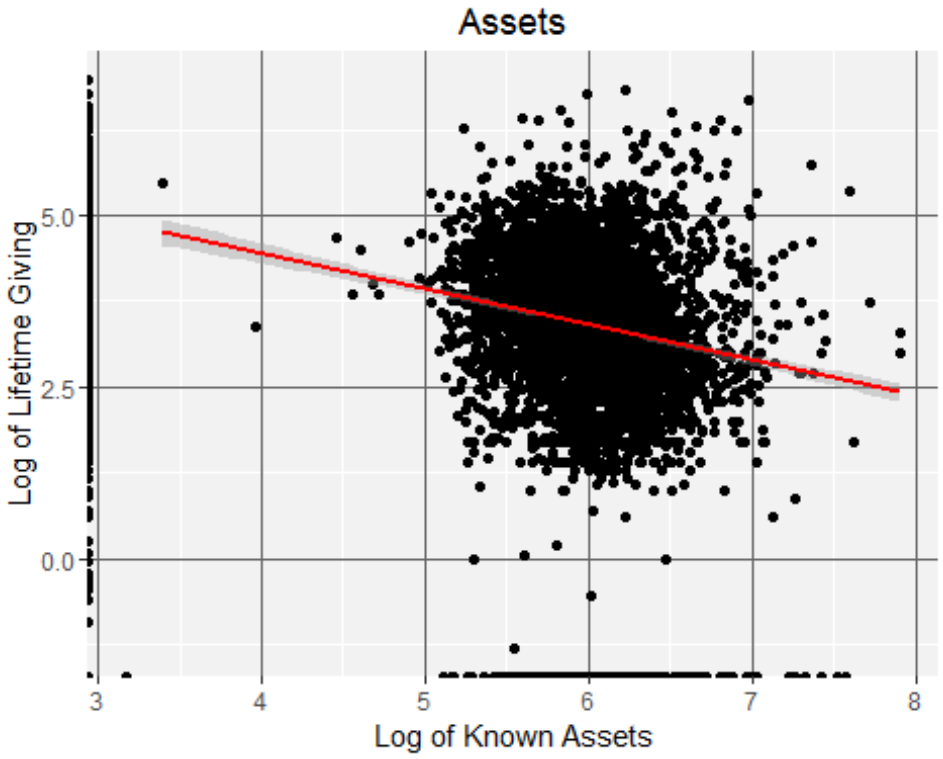
The variable Number.of.Relationships will be included in the regression model since it seems to have a relationship to giving.

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?

Assets

As with Lifetime Giving, the Asset variable was log transformed.

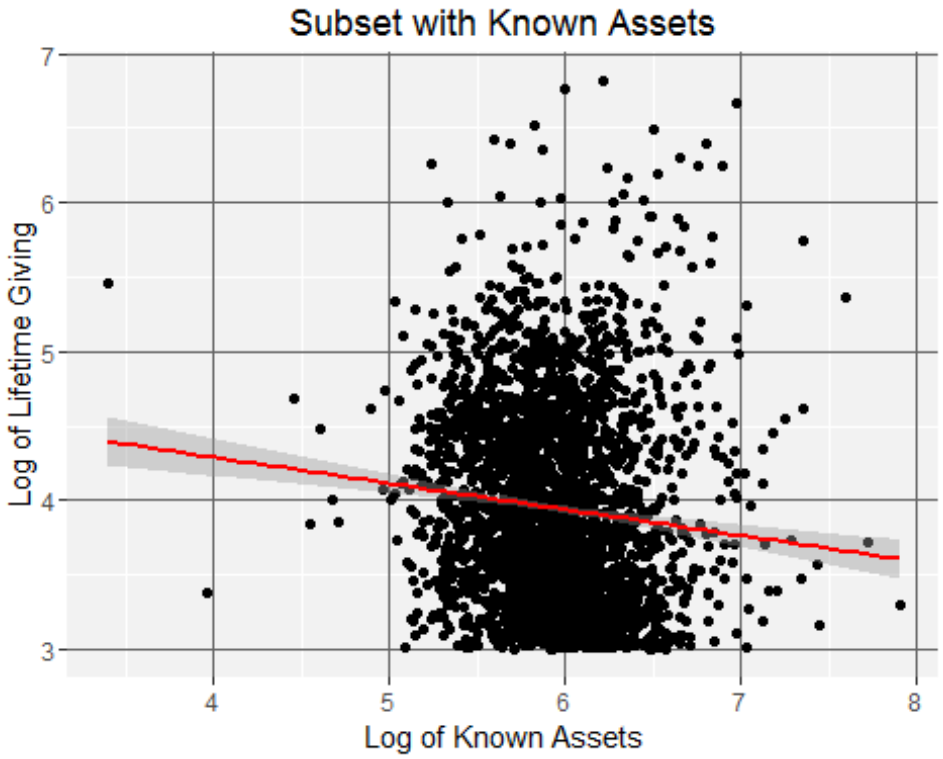
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    -Inf    -Inf   5.694    -Inf   6.112   7.903       1
```



There are a couple of interesting things going on in this plot. The large mass in the center of the graph shows the expected relationship between having assets and lifetime giving--those who have often give. The solid line on the left of giving by those with no known assets could be a data collection issue, since asset data has not been collected for people who have given small amounts.

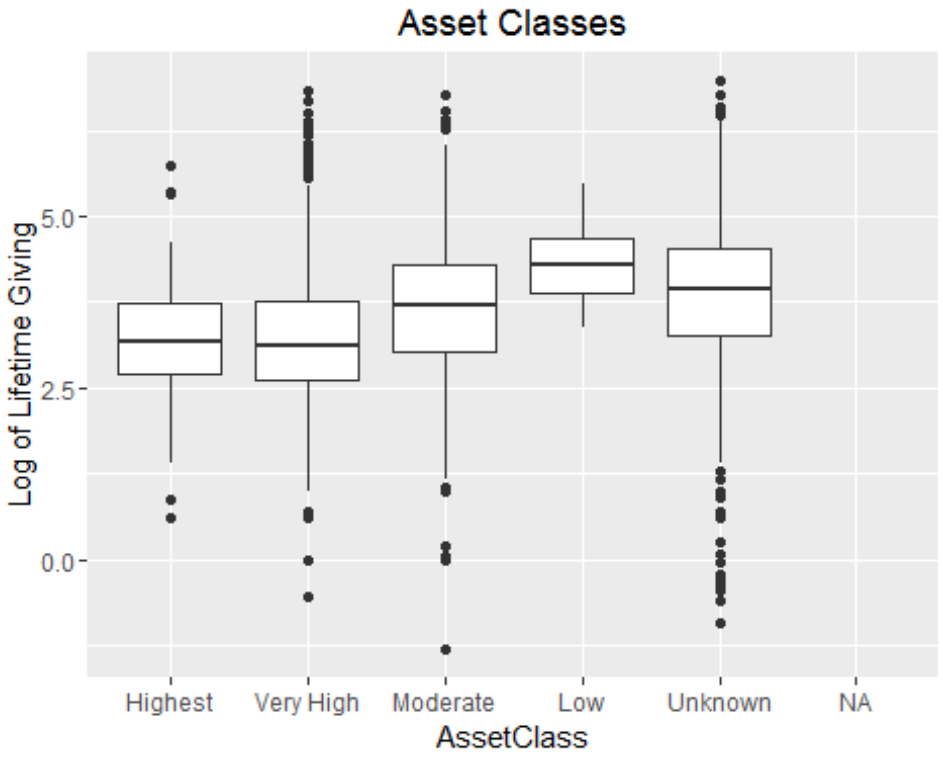
The regression line shows a negative relationship. Since this relationship might be influenced by $0 in lifetime giving and assets, the data will be subsetted to exclude them and plotted again.

```
TS4 <- subset(TS, ASSETS > 0 & WSU.LIFETIME.GIVING > 1000)
subassets <- ggplot(TS4, aes(x=logAssets,y=logGiving))+geom_jitter()+ stat_sm
ooth(method="lm", col="red")+ theme(panel.background = element_rect(fill = "g
rey95"),panel.grid.major = element_line(colour = "grey40"))
subassets <- subassets + scale_x_continuous(name="Log of Known Assets") + sca
le_y_continuous(name = "Log of Lifetime Giving")
subassets <- subassets + ggtitle("Subset with Known Assets")
subassets
```



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

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), this box plot was made:

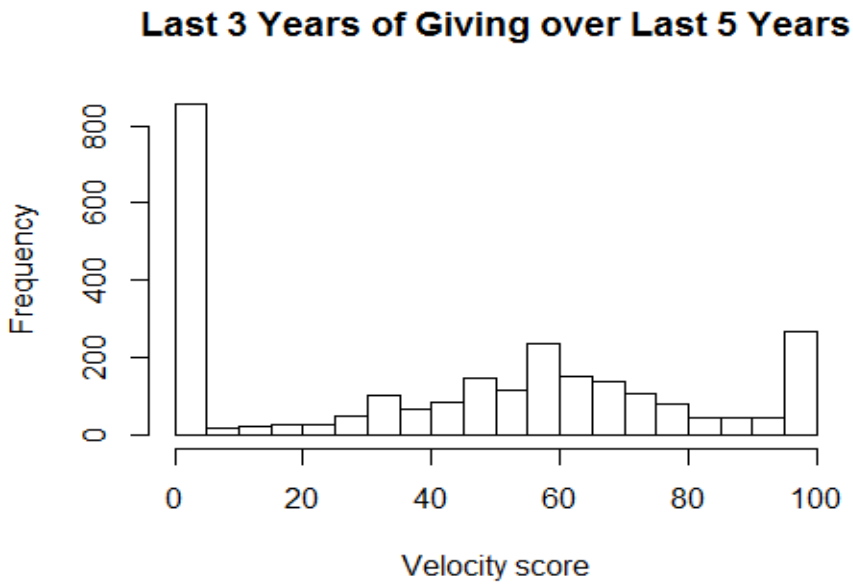


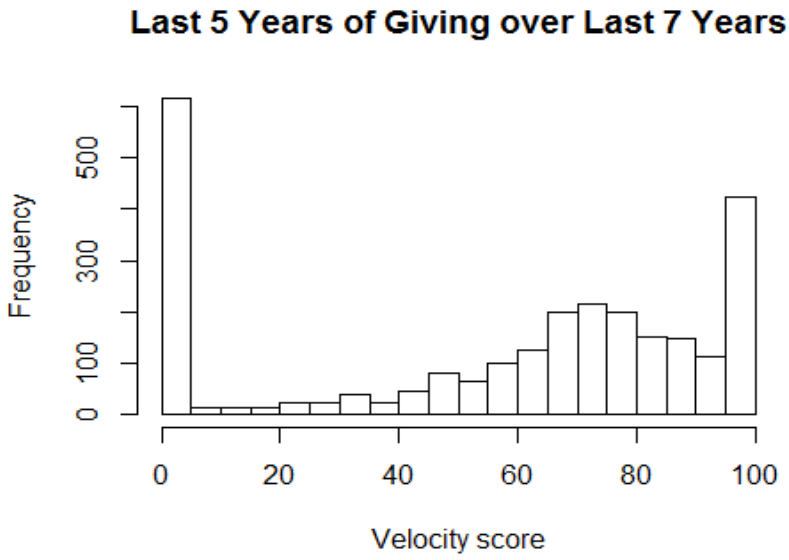
This box plot shows the expected relationships between having assets and making donations in the "Highest", "Very High" and "Moderate" classes. For this reason, the variable ASSETS will be included in the regression model.

Of more interest 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. Since the easiest asset data to find is the value of real estate, business 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.

Velocity and Lifetime Giving

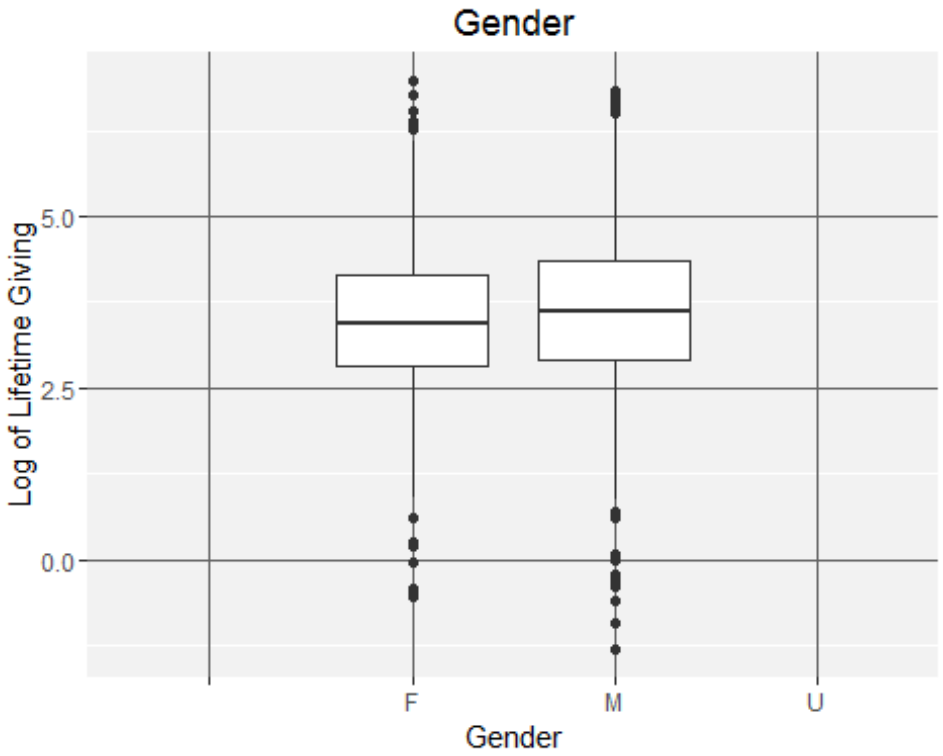
Velocity is a measure of recent giving. For the purposes of this study, I have calculated it 2 different ways and graph them against Lifetime Giving.





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. For later modeling purposes, the first Velocity Score (last 3 years of giving divided by last 5 years of giving) will be included in the regression model since it more closely resembles a normal distribution.

Gender

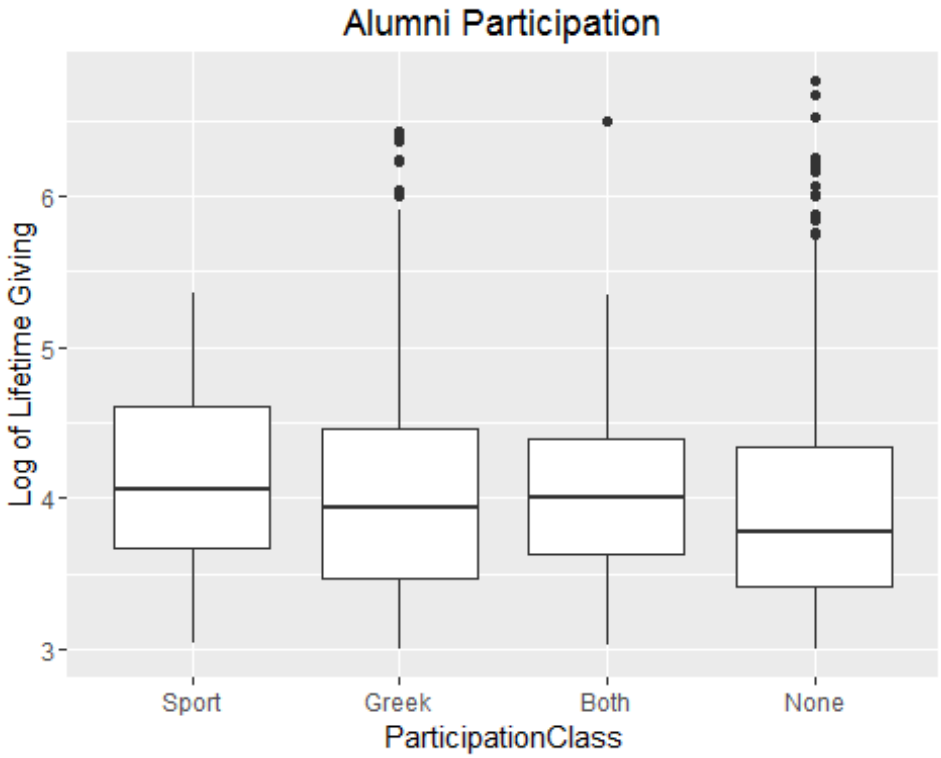


This box plot does not show a great difference in giving related to gender. Since it is unlikely to have any predictive value for giving, the variable GENDER will not be included in the regression model.

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.

```
TS6 <-subset(TS, ASSETS > 0 & WSU.LIFETIME.GIVING > 1000 & AlumniCode > 0)
```



This score is computed by assigning one "point" for membership in a Greek chapter or sports club, then totaling the points. It looks like participation in either a fraternity or sorority has a more favorable relationship to lifetime giving than participation in a sport, but the "None" seems influential as well.

Welch's Two Sample t-test was performed on these variables.

```
Spt = TS$SportCode == 1
PlayedSport=TS[Spt,]$WSU.LIFETIME.GIVING  #Lifetime Giving by those who playe
d Sports

NSpt = TS$SportCode ==0
NoSport=TS[NSpt,]$WSU.LIFETIME.GIVING #Lifetime Giving by those who didn't pl
ay Sports

t.test(PlayedSport, NoSport)

##
##  Welch Two Sample t-test
##
```

```
## data:  PlayedSport and NoSport
## t = 0.92664, df = 415.35, p-value = 0.3547
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12330.27  34322.91
## sample estimates:
## mean of x mean of y
##  51079.61  40083.30
```

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

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

t.test(Greek, GDI)
```

```
##
##  Welch Two Sample t-test
##
## data:  Greek and GDI
## t = 2.7324, df = 3633.6, p-value = 0.006317
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4803.815 29209.996
## sample estimates:
## mean of x mean of y
##  53514.57  36507.66
```

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

```
##
##  Welch Two Sample t-test
##
## data:  PlayedSport and Greek
## t = -0.19418, df = 513.87, p-value = 0.8461
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27069.76  22199.86
## sample estimates:
## mean of x mean of y
##  51079.61  53514.57
```

Given the differences in the means of these samples, GreekCode and SportCode should be part of the regression model.

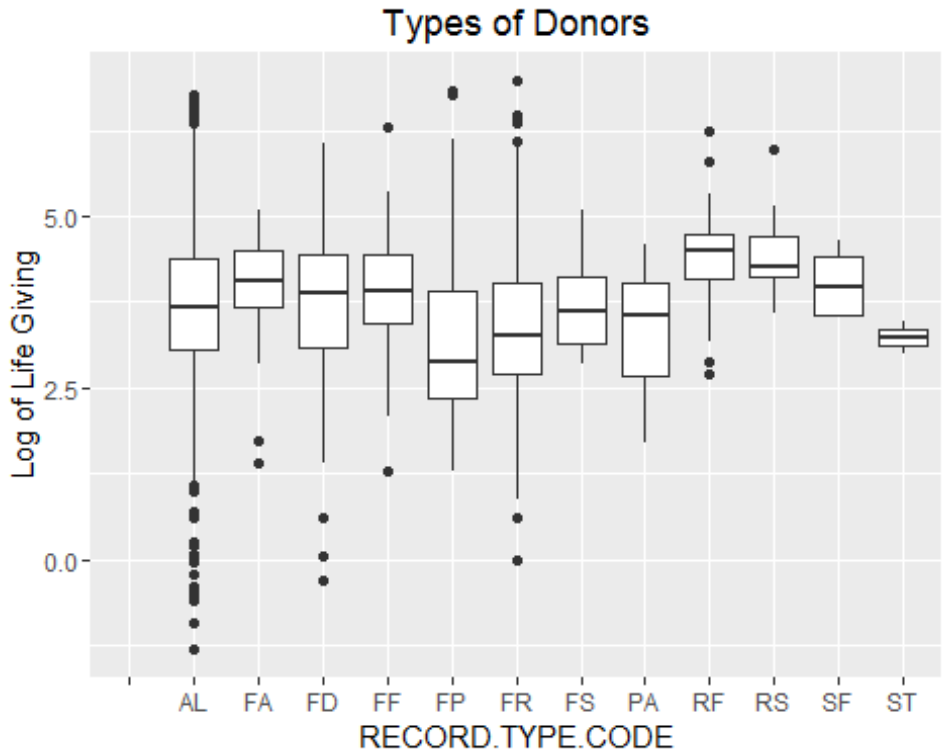
Types of donors in the sample

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.

```
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 since it has a large, potentially significant, effect on giving.

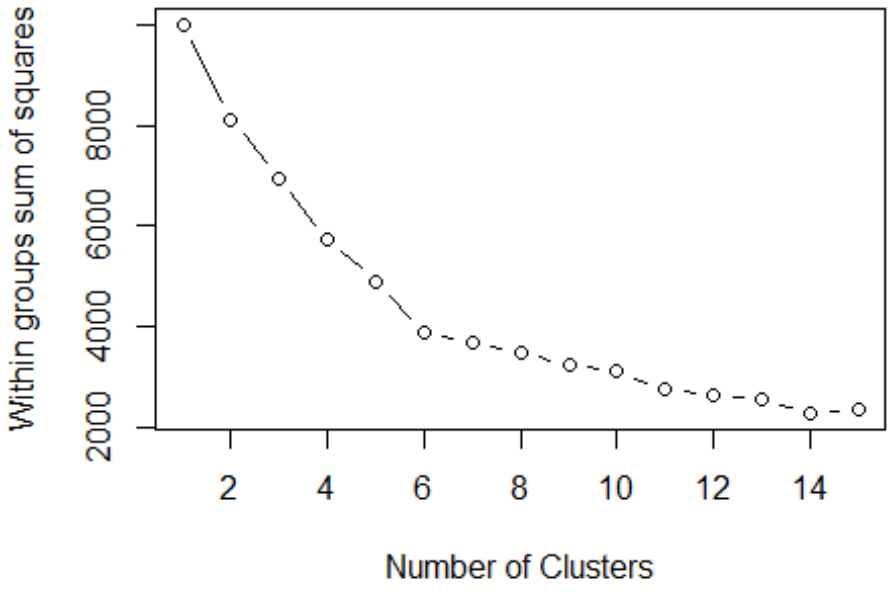
K Means Clustering

This attempt at clustering, uses a subset of the TS6 dataset and includes the WSU.LIFETIME.GIVING, WSU.YEARS.OF.GIVING, ASSETS, Number.of.relationships, and Velocity57Score variables. The resulting Donor dataset includes 2005 observations.

```
Donor <- TS6[c(2,3,4,6,8)]
Donor <- na.omit(Donor)
Donor <- scale(Donor)
```

Determine the number of clusters:

```
wss <- (nrow(Donor)-1)*sum(apply(Donor,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(Donor, centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum
of squares")
```



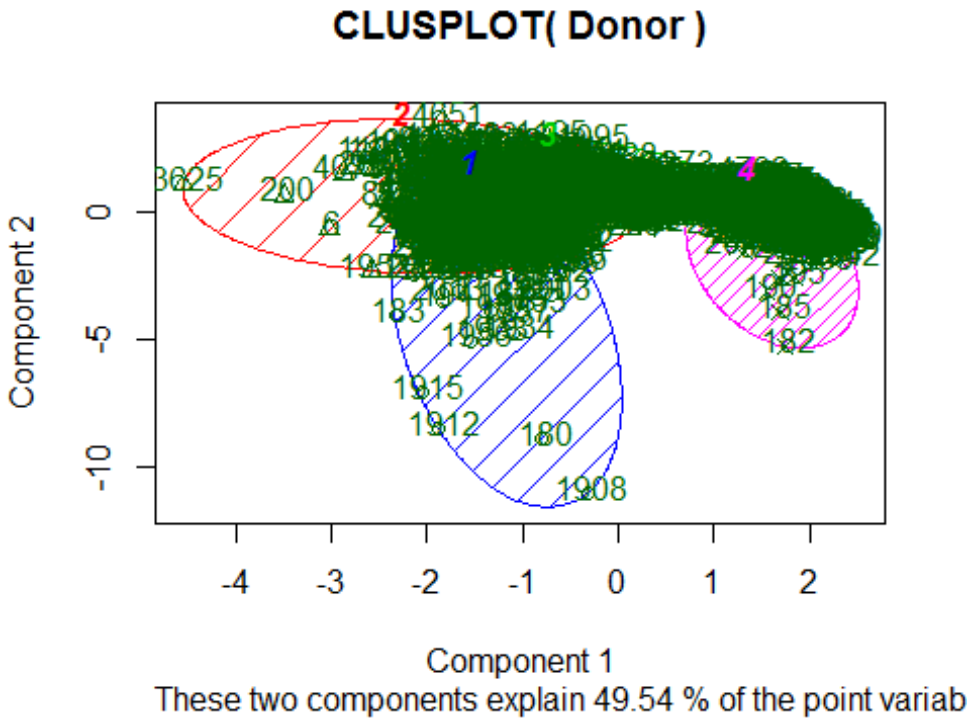
This graph indicates 4 clusters should be used for the analysis.

Run k means cluster analysis:

```
fit <- kmeans(Donor, 4)
aggregate(Donor,by=list(fit$cluster),FUN=mean)

##   Group.1 WSU.LIFETIME.GIVING WSU.YEARS.OF.GIVING      ASSETS
## 1       1         -0.08362941          -0.5612542  0.27788955
## 2       2          0.62594430           0.3010377 -0.03891499
## 3       3         -0.02185190           0.9773563 -0.23870688
## 4       4         -0.15141779          -0.6800478  0.03722791
##   Number.of.relationships Velocity35Score
## 1              -0.2997858       0.9077919
## 2               2.1020484       0.2224282
## 3              -0.2425485       0.3296034
## 4              -0.3158344      -1.1756323

Donor  <- data.frame(Donor, fit$cluster)
clusplot(Donor, fit$cluster, color=TRUE, shade=TRUE,labels=2, lines=0)
```



Due to the overlapping nature of the clusters and the density of points in the clusters, it is difficult to visually see what is going on here. This might be the result of no variables in the subset being significant enough in relation to the others to make the clusters more distinct.

Regression Model

When these 7 variables are included (WSU.YEARS.OF.GIVING, ASSETS, Number.of.relationships, AlumniCode, SportCode, GreekCode, Velocity35Score) in the 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 and these variable can be said to have a significant relationship to Lifetime Giving.

```
RegMod <-lm(WSU.LIFETIME.GIVING~WSU.YEARS.OF.GIVING+ASSETS+Number.of.relation
ships+AlumniCode+SportCode+GreekCode+Velocity35Score,data=TS)
summary(RegMod)

##
## Call:
## lm(formula = WSU.LIFETIME.GIVING ~ WSU.YEARS.OF.GIVING + ASSETS +
##     Number.of.relationships + AlumniCode + SportCode + GreekCode +
##     Velocity35Score, data = TS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -236999  -47052  -22297     901 9520790
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.427e+03  6.640e+03  -0.667    0.505
## WSU.YEARS.OF.GIVING       2.880e+03  3.029e+02   9.507  < 2e-16 ***
## ASSETS                    1.840e-03  1.288e-03   1.428    0.153
## Number.of.relationships   1.367e+04  2.068e+03   6.608 4.21e-11 ***
## AlumniCode               -3.729e+04  7.874e+03  -4.737 2.22e-06 ***
## SportCode                 3.238e+03  1.452e+04   0.223    0.824
## GreekCode                 4.809e+02  7.815e+03   0.062    0.951
## Velocity35Score           6.643e+01  9.603e+01   0.692    0.489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248900 on 6472 degrees of freedom
##   (837 observations deleted due to missingness)
## Multiple R-squared:  0.02934,    Adjusted R-squared:  0.02829
## F-statistic: 27.94 on 7 and 6472 DF,  p-value: < 2.2e-16
```

Running the regression model again, but this time using the smaller TS6 dataset, results in a p-value still below 0.05. ASSETS, Number.of.relationships,and to a lesser degree, WSU.YEARS.OF.GIVING are shown to be significant.

```
RegModTS6 <-lm(WSU.LIFETIME.GIVING~WSU.YEARS.OF.GIVING+ASSETS+Number.of.relat
ionships+AlumniCode+SportCode+GreekCode+Velocity35Score, data = TS6)
summary(RegModTS6)

##
## Call:
## lm(formula = WSU.LIFETIME.GIVING ~ WSU.YEARS.OF.GIVING + ASSETS +
##     Number.of.relationships + AlumniCode + SportCode + GreekCode +
##     Velocity35Score, data = TS6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -450535  -47740  -25938   -3103 5808956
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.125e+04  1.596e+04  -2.584  0.00984 **
## WSU.YEARS.OF.GIVING       1.493e+03  5.247e+02   2.846  0.00447 **
## ASSETS                    1.710e-02  3.548e-03   4.819 1.55e-06 ***
## Number.of.relationships   1.365e+04  3.370e+03   4.052 5.28e-05 ***
## AlumniCode                       NA         NA      NA       NA
## SportCode                -2.470e+03  2.112e+04  -0.117  0.90694
## GreekCode                -5.725e+03  1.145e+04  -0.500  0.61724
## Velocity35Score           2.722e+02  1.686e+02   1.615  0.10656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 247500 on 1998 degrees of freedom
##   (100 observations deleted due to missingness)
## Multiple R-squared:  0.02616,    Adjusted R-squared:  0.02323
## F-statistic: 8.945 on 6 and 1998 DF,  p-value: 1.166e-09
```

The low R-squared value produced by both models indicate considerable variability in the data. Some variables are significant predictors, but the range of the prediction is very large. But most variables show a high p-value, meaning they are not predictive of giving.

Project Conclusions

The graphs of most of the variables in the TS dataset pointed to some correlation with lifetime giving, but didn't give any indication of their significance.

The attempt at k means clustering didn't really produce any meaningful clusters. This is probably a result of the variablity in the data.

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. Neither model is very good at dealing with the variability of the data in the large TS sample, or the somewhat more restricted TS6 sample.

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.

Future Directions

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 a Largest Gift Score might be significant predictors. A calculation of how many miles away a donor lives, or how old the donor was when they gave their first gift would also be interesting to examine.

It would be very interesting to look at covariance and if any of these variable are working together to influence giving. Is distance a predictive variable simply because major donors live in expensive neighborhoods? There might be variables that turn out to be proxies for data points that are not represented in the donor database. This study has been only a starting point in examining and understanding data that can be extracted from the donor database.