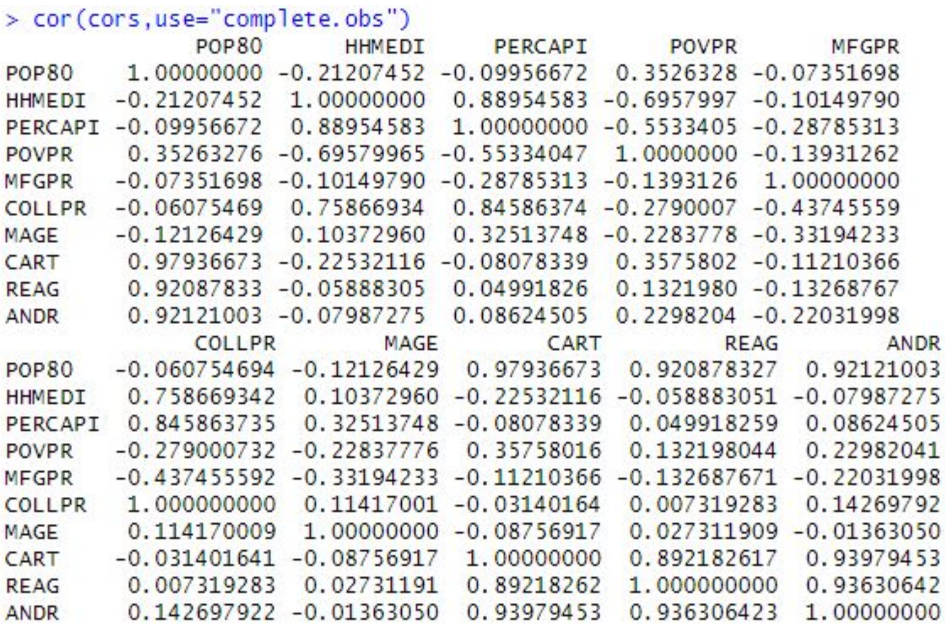
ISOM-670 Business Statistics

SeaWatch Assignment

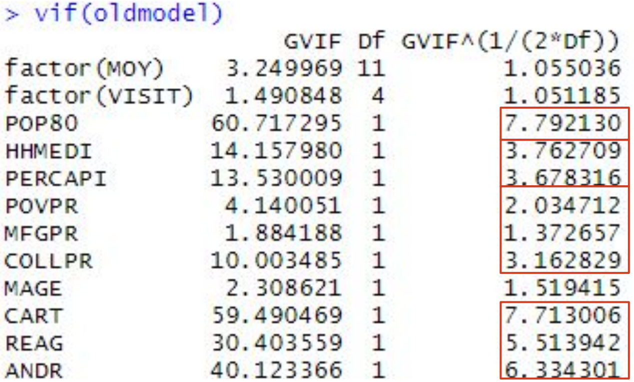
Team 2: Yifei, Jie, Eleanor & Susie

Full Model Approach

- Full model linear regression (all variables except MON & YR with factor(MOY) and factor(VISIT).
- Good:
  - High R-squared
  - most variables have low p-value. ⥀
- Bad:
  - Extremely high VIFs, indicating high multicolinearity between variables
  - High correlation between:
    - POP80 & (CART, REAG, ANDR)
    - (HHMEDI, PERCAPI) & (POVPR, COLLPR)
- We need a new approach to lower multicollinearity by excluding variables and creating new variables.



```
> cor(cors,use="complete.obs")
              POP80      HHMEDI     PERCAPI      POVPR       MFGPR
POP80    1.00000000 -0.21207452 -0.09956672  0.3526328 -0.07351698
HHMEDI  -0.21207452  1.00000000  0.88954583 -0.6957997 -0.10149790
PERCAPI -0.09956672  0.88954583  1.00000000 -0.5533405 -0.28785313
POVPR    0.35263276 -0.69579965 -0.55334047  1.0000000 -0.13931262
MFGPR   -0.07351698 -0.10149790 -0.28785313 -0.1393126  1.00000000
COLLPR  -0.06075469  0.75866934  0.84586374 -0.2790007 -0.43745559
MAGE    -0.12126429  0.10372960  0.32513748 -0.2283778 -0.33194233
CART     0.97936673 -0.22532116 -0.08078339  0.3575802 -0.11210366
REAG     0.92087833 -0.05888305  0.04991826  0.1321980 -0.13268767
ANDR     0.92121003 -0.07987275  0.08624505  0.2298204 -0.22031998
              COLLPR        MAGE        CART         REAG        ANDR
POP80   -0.060754694 -0.12126429  0.97936673  0.920878327  0.92121003
HHMEDI   0.758669342  0.10372960 -0.22532116 -0.058883051 -0.07987275
PERCAPI  0.845863735  0.32513748 -0.08078339  0.049918259  0.08624505
POVPR   -0.279000732 -0.22837776  0.35758016  0.132198044  0.22982041
MFGPR   -0.437455592 -0.33194233 -0.11210366 -0.132687671 -0.22031998
COLLPR   1.000000000  0.11417001 -0.03140164  0.007319283  0.14269792
MAGE     0.114170009  1.00000000 -0.08756917  0.027311909 -0.01363050
CART    -0.031401641 -0.08756917  1.00000000  0.892182617  0.93979453
REAG     0.007319283  0.02731191  0.89218262  1.000000000  0.93630642
ANDR     0.142697922 -0.01363050  0.93979453  0.936306423  1.00000000
```



```
> vif(oldmodel)
                     GVIF Df GVIF^(1/(2*Df))
factor(MOY)      3.249969 11        1.055036
factor(VISIT)    1.490848  4        1.051185
POP80           60.717295  1        7.792130
HHMEDI          14.157980  1        3.762709
PERCAPI         13.530009  1        3.678316
POVPR            4.140051  1        2.034712
MFGPR            1.884188  1        1.372657
COLLPR          10.003485  1        3.162829
MAGE             2.308621  1        1.519415
CART            59.490469  1        7.713006
REAG            30.403559  1        5.513942
ANDR            40.123366  1        6.334301
```

⥀ See **Appendix** for the summary of linear regression for the full model approach.

Our Approach

- Donation Received can be explained by:

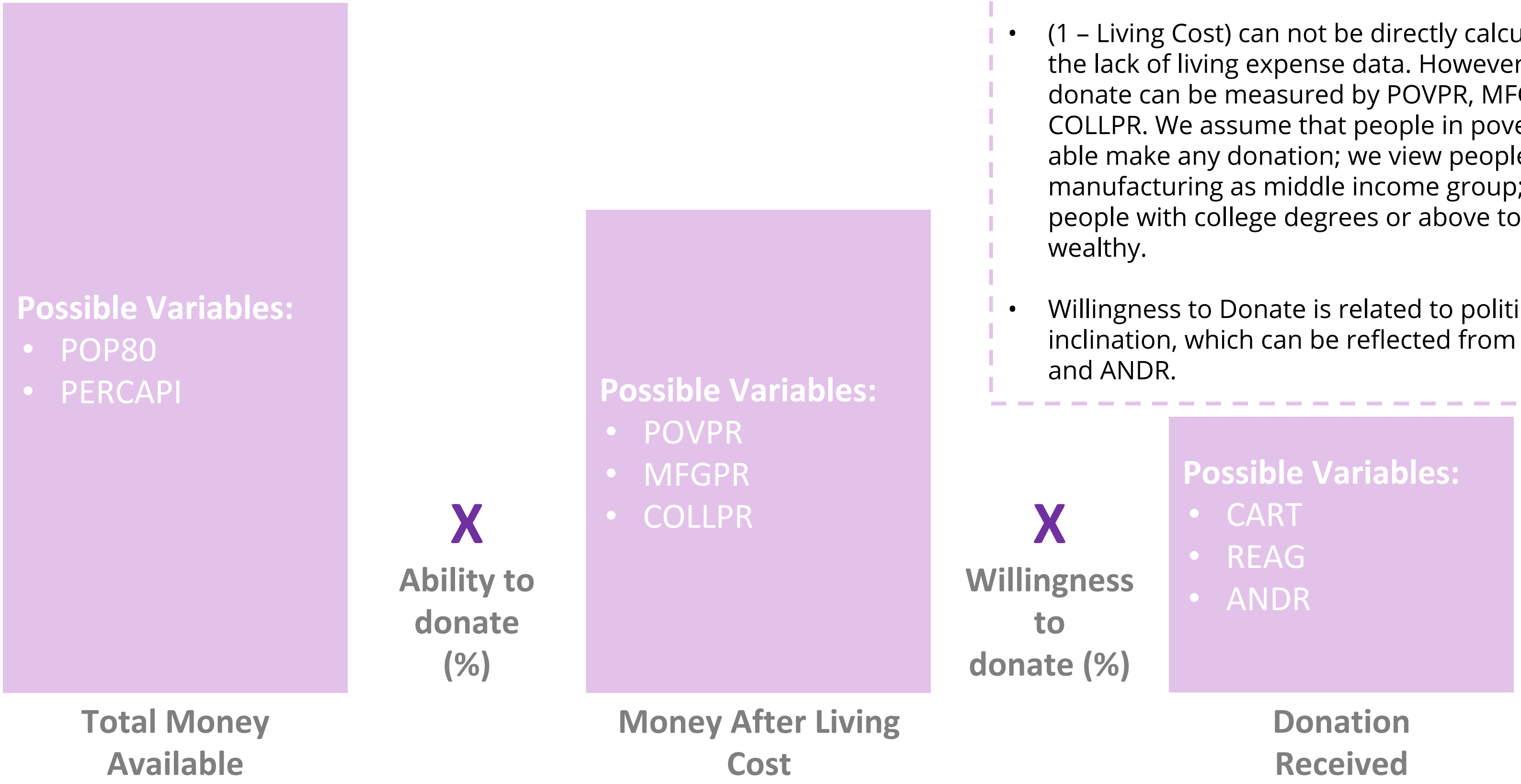
Donation = Total Money Available * (1 - Living Cost) * Willingness to Donate(% of rest income)

- Total Money Available can be calculated from:

Money Available = Population * Income Per Person

- (1 – Living Cost) can not be directly calculated, due to the lack of living expense data. However, the ability to donate can be measured by POVPR, MFGPR, and COLLPR. We assume that people in poverty won't be able make any donation; we view people in manufacturing as middle income group; we consider people with college degrees or above to be more wealthy.

- Willingness to Donate is related to political inclination, which can be reflected from CART, REAG, and ANDR.



Possible Variables:
- POP80
- PERCAPI

Total Money Available

X

Ability to donate (%)

Possible Variables:
- POVPR
- MFGPR
- COLLPR

Money After Living Cost

X

Willingness to donate (%)

Possible Variables:
- CART
- REAG
- ANDR

Donation Received

New Variables

(1) Total Money Available

(a) Standardized Total Gross Receipts (DGROSS) : to reveal the real monetary value of GROSS in each period, we use the latest CPI as the base point
> DGROSS <- GROSS / CPI * 300.9 (most recent CPI in 1983 August)

(b) Total Money Available (TMAV) : to show how many money are possibly available in each town, we create this index
> TMAV <- POP80 * PERCAPI

(2) Money After Living Cost

(a) Wealth Index (WI): to quantify and measure different classes' the abilities to donate after deducting essential living cost. For coefficients, we use 0 for POVPR, use 1 for MFGPR, and 1.4 for COLLPR. ⥀
> WI <- 0*POVPR + 1*MFGPR + 1.4*COLLPR

(3) Willingness to Donate

(a) Total Votes (VOTE): to calculate the total votes of a town because not everyone is voting
> VOTE <- CART + REAG + ANDR

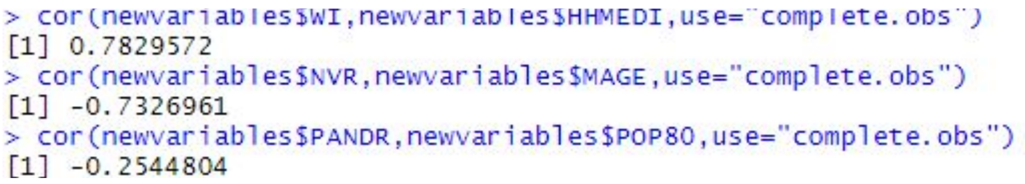
(b) Non-voting Rate (NVR): to represent those people who have no preference in voting, in other words, no strong willingness to donate
> NVR <- (POP80 - VOTE) / POP80

(c) Anderson's Percentage in Total Votes (PANDR): we assume people who support Anderson tend to be more liberal, this they are more concerned with environmental issue so we take ANDR into consideration
> PANDR <- ANDR/VOTE

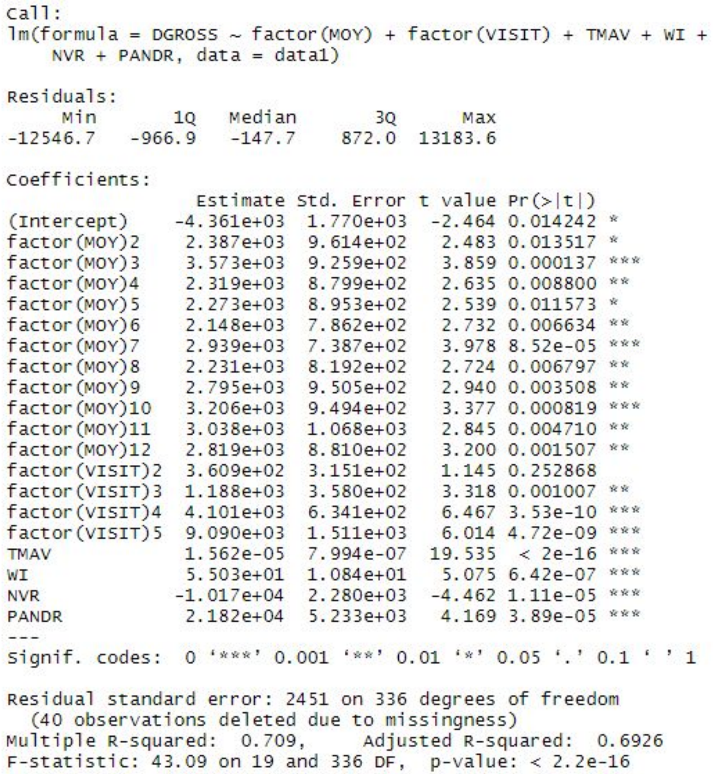
⥀ In 1980, college degree holders have a 40% income premium compared to people with only high school degrees. Figure 8 on Pg.27 of http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.197.9102&rep=rep1&type=pdf

Linear Regression: Final Model Used

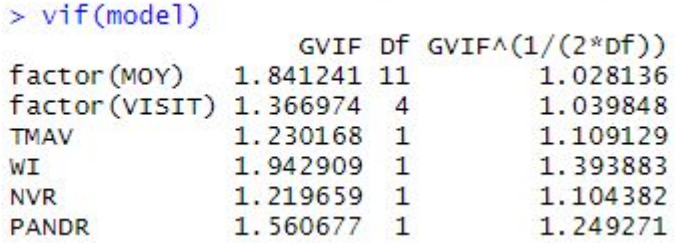
- Our final model includes factor(MOY), factor(VISIT), TMAV, WI, NVR, PANDR
- R-squared lowers to 0.709, still a good number
- Good:
  - All variables are statistically significant
  - VIFs decrease to around 1; max VIF is 1.39 compared to 7.79 before
  - WI does a good job measuring the level of wealthiness (high correlation between WI & HHMEDI)
  - NVR accidentally captures the effect of MAGE (young people generally don’t care much about politics)
  - PANDR(ANDR/VOTE) doesn’t have much correlation with POP80, while still successfully measuring political inclination



```
> cor(newvariables$WI,newvariables$HHMEDI,use="complete.obs")
[1] 0.7829572
> cor(newvariables$NVR,newvariables$MAGE,use="complete.obs")
[1] -0.7326961
> cor(newvariables$PANDR,newvariables$POP80,use="complete.obs")
[1] -0.2544804
```



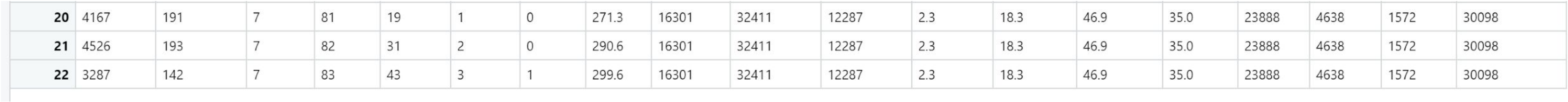
```
Call:
lm(formula = DGROSS ~ factor(MOY) + factor(VISIT) + TMAV + WI + 
    NVR + PANDR, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max 
-12546.7   -966.9   -147.7    872.0  13183.6 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -4.361e+03  1.770e+03  -2.464 0.014242 *  
factor(MOY)2    2.387e+03  9.614e+02   2.483 0.013517 *  
factor(MOY)3    3.573e+03  9.259e+02   3.859 0.000137 ***
factor(MOY)4    2.319e+03  8.799e+02   2.635 0.008800 ** 
factor(MOY)5    2.273e+03  8.953e+02   2.539 0.011573 *  
factor(MOY)6    2.148e+03  7.862e+02   2.732 0.006634 ** 
factor(MOY)7    2.939e+03  7.387e+02   3.978 8.52e-05 ***
factor(MOY)8    2.231e+03  8.192e+02   2.724 0.006797 ** 
factor(MOY)9    2.795e+03  9.505e+02   2.940 0.003508 ** 
factor(MOY)10   3.206e+03  9.494e+02   3.377 0.000819 ***
factor(MOY)11   3.038e+03  1.068e+03   2.845 0.004710 ** 
factor(MOY)12   2.819e+03  8.810e+02   3.200 0.001507 ** 
factor(VISIT)2  3.609e+02  3.151e+02   1.145 0.252868    
factor(VISIT)3  1.188e+03  3.580e+02   3.318 0.001007 ** 
factor(VISIT)4  4.101e+03  6.341e+02   6.467 3.53e-10 ***
factor(VISIT)5  9.090e+03  1.511e+03   6.014 4.72e-09 ***
TMAV            1.562e-05  7.994e-07  19.535  < 2e-16 ***
WI              5.503e+01  1.084e+01   5.075 6.42e-07 ***
NVR            -1.017e+04  2.280e+03  -4.462 1.11e-05 ***
PANDR           2.182e+04  5.233e+03   4.169 3.89e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2451 on 336 degrees of freedom
  (40 observations deleted due to missingness)
Multiple R-squared:  0.709,	Adjusted R-squared:  0.6926 
F-statistic: 43.09 on 19 and 336 DF,  p-value: < 2.2e-16
```



```
> vif(model)
                  GVIF Df GVIF^(1/(2*Df))
factor(MOY)   1.841241 11        1.028136
factor(VISIT) 1.366974  4        1.039848
TMAV          1.230168  1        1.109129
WI            1.942909  1        1.393883
NVR           1.219659  1        1.104382
PANDR         1.560677  1        1.249271
```

Cleaning Data and Modify Inappropriate Data

We reviewed each variable in histogram to find inappropriate data.
(1) In data C, we observe that in two cities (six datasets)- Agawam, Longmeadow- total number of voters are higher than population.
Therefore, we turn “votes” for the three candidates into NA.



20	4167	191	7	81	19	1	0	271.3	16301	32411	12287	2.3	18.3	46.9	35.0	23888	4638	1572	30098
21	4526	193	7	82	31	2	0	290.6	16301	32411	12287	2.3	18.3	46.9	35.0	23888	4638	1572	30098
22	3287	142	7	83	43	3	1	299.6	16301	32411	12287	2.3	18.3	46.9	35.0	23888	4638	1572	30098

(2) In data D, we find there are many missing values in the variables used in our model(e.g.: some towns’ “HHMEDI” & “PERCAPI” are NAs, while others have many missing political data), so GROSS prediction for these data points would be impossible. We study the four groups of cities based on their proximity to Greenwich and Bridgeport -- “GRN=1”(within 25 mi. of Greenwich), “GRN=2”(25-40 mi. of Greenwich), “BPT=1”(within 25 mi. of Bridgeport), “BPT=2”(25-40 mi. of Bridgeport). After predicting with our model, we notice that the situations where prediction is impossible are actually few relative to the total counts within each group, thus not impacting our result very much. We immute the missing data of these variables with their median in the end.

(3) We decide to exclude four data points from New York City because the board has been convinced to skip NYC for the foreseeable future.

Prediction Result: SeaWatch (D) Data

Counts by Proximity to two office site choices

	Greenwich	Bridgeport
Number of cities within approx. 25 miles	49	50
Number of cities locate 25-40 miles away	51	66

For further analysis, we use index (1-4) to represent each cities group as below:

	Greenwich	Bridgeport
< 25 miles	1	3
25-40 miles	2	4

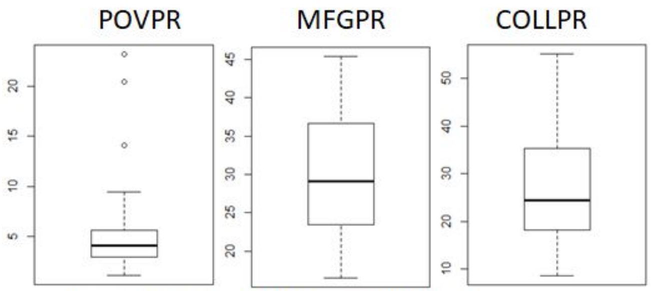
Predicted DGROSS by Proximity to two office site choices

	Greenwich	Bridgeport
Number of cities within approx. 25 miles	$ 290737	$ 191602.6
Number of cities locate 25-40 miles away	$ 303769.9	$ 173233.2
SUM	**$ 594506.9**	**$ 364835.8**

Conclusion: Greenwich is more worthy for Sea Watch to go after because cities closer to Greenwich are highly likely to receive more donation. In addition, the underlying reasons we found to cause this outcome is due to political inclination rather than demographics (like age and population) or wealthiness of a town (like money available and willingness to donate)

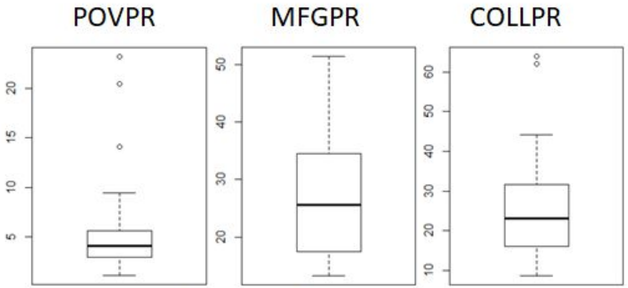
Social Class Distribution for Four Sites Group

[3rd] BPT1 - diverse



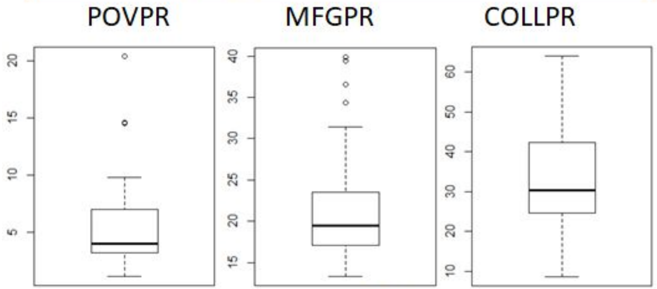
> social class between each cities has larger differences

[4th] BPT2 – balanced



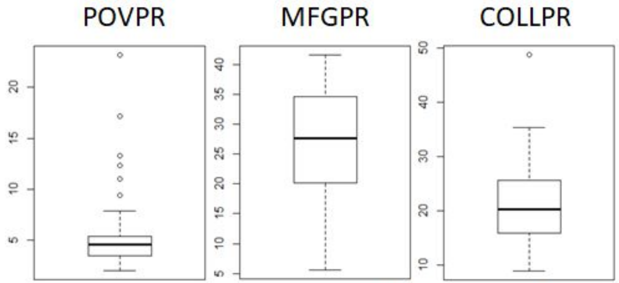
> The distribution of class among different cities is more balanced

[2nd] GRN1 – social elites



> Most cities have more social elites with higher education

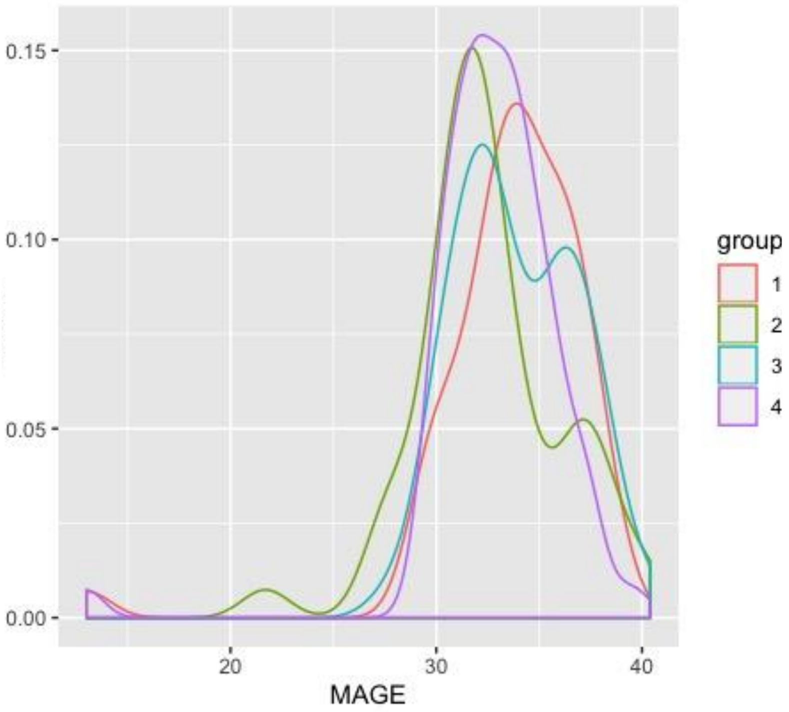
[1st] GRN2 – middle class



> Most cities have more middle class people who works in manufacturing industry

As mentioned before, we regard: people live beneath poverty level as the poor, people work in manufacturing industry as the middle class, people has attained higher education than four-year college as the rich (social elites).
From the analysis on the left, we could see class differences could not explicitly explain the potential donation we could received. For example, GRN2 has the highest potential donation but middle class was the majority in most cities. On the other hand, GRN1 that has more people having higher education goes to the second place.

Distribution of “POP80” & “MAGE” for Four Site Groups

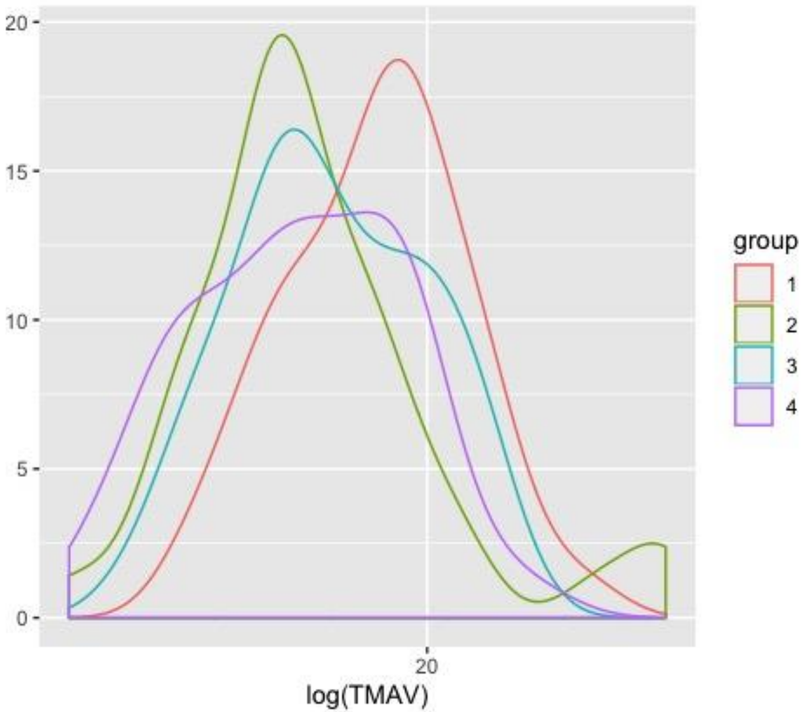


From the graph and statistics, we can see that towns near Greenwich generally have a much bigger population than Bridgeport, and according to our regression result, Greenwich may have a higher potential for donation than Bridgeport. So a big population could be an important factor for donation to be received.

As for “Age”, towns within 25-40 miles from Greenwich are populated with younger people, while towns within 25 miles from Greenwich are older. Age does not seem to affect donation a lot.

	mean(POP80)	**mean(MAGE)**
“GRN=1” group	41347	33.68
“GRN=2” group	44570	32.54
“BPT=1” group	28704	33.89
“BPT=2” group	22824	32.86

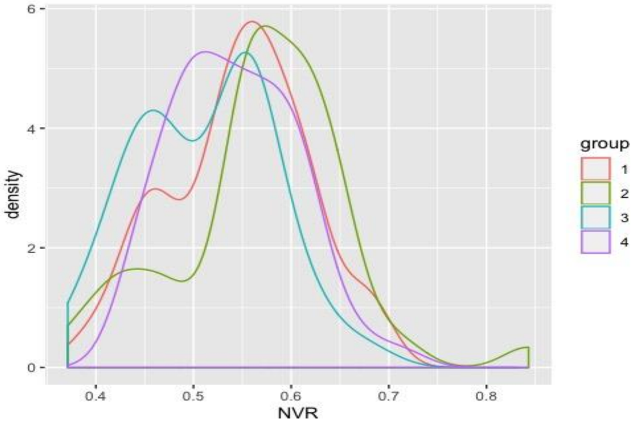
Distribution of “TMAV” & “WI” for Four Site Groups

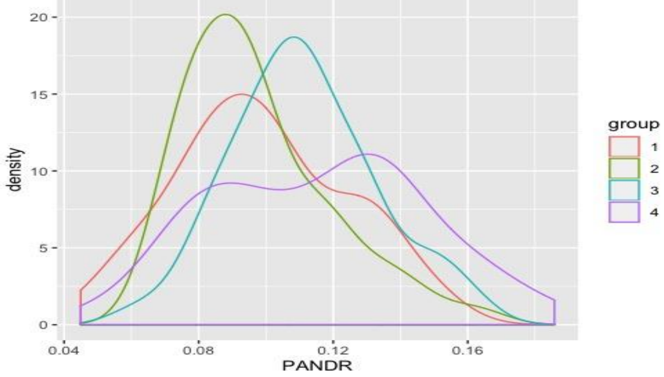


From the graph and statistics, it’s clear that the total money availability of towns near Greenwich is generally higher, which may lead to a greater donation to be received. Therefore, the total money availability should indeed be an important contributing factor for the amount of donation.

	mean(TMAV)	**mean(WI)**
“GRN=1” group	437666408	67.4
“GRN=2” group	522075215	57.9
“BPT=1” group	283995754	67.7
“BPT=2” group	225009862	61.4

Distribution of “NVR” & “PANDR” for Four Site Groups

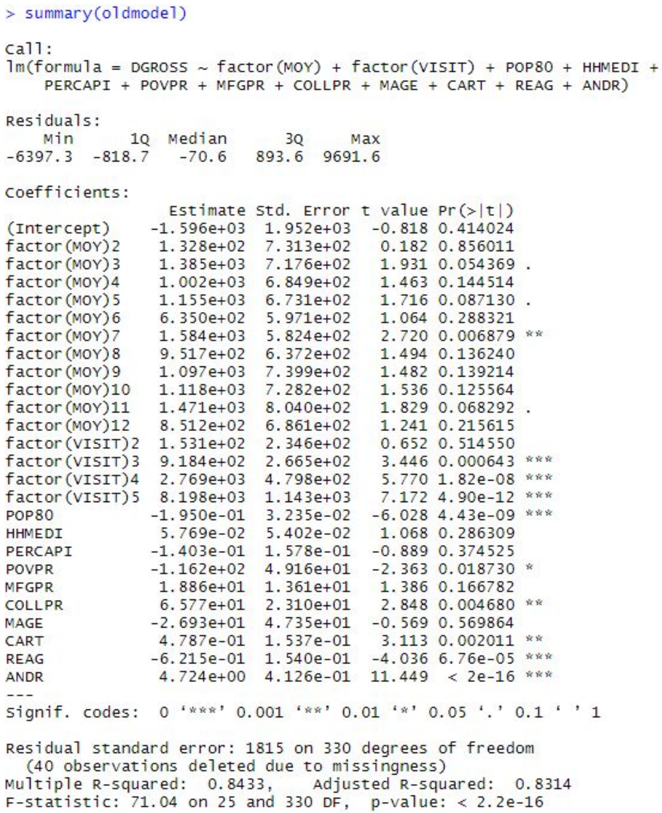




From the graph and statistics, we have an interesting observation in particular that political inclination might be among the most important factor for donation to be received. Towns within 25 miles from Greenwich have the highest rate of supporting Anderson, indicating pro-environmental inclination; those within 25-40 miles from Greenwich have the greatest non-voter rate, which may indicate less interest in politics. We may conclude that the inclination for voting for Anderson could signal a higher amount of donation to obtain..

	mean(NVR)	**mean(PANDR)**
“GRN=1”	0.5465	0.0985
“GRN=2”	0.5697	0.0973
“BPT=1”	0.5111	0.1111
“BPT=2”	0.5416	0.1154

Appendix: Linear Regression - Full Model Approach



```
> summary(oldmodel)

Call:
lm(formula = DGROSS ~ factor(MOY) + factor(VISIT) + POP80 + HHMEDI +
    PERCAPI + POVPR + MFGPR + COLLPR + MAGE + CART + REAG + ANDR)

Residuals:
    Min      1Q  Median      3Q     Max
-6397.3  -818.7   -70.6   893.6  9691.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.596e+03  1.952e+03  -0.818 0.414024
factor(MOY)2    1.328e+02  7.313e+02   0.182 0.856011
factor(MOY)3    1.385e+03  7.176e+02   1.931 0.054369 .
factor(MOY)4    1.002e+03  6.849e+02   1.463 0.144514
factor(MOY)5    1.155e+03  6.731e+02   1.716 0.087130 .
factor(MOY)6    6.350e+02  5.971e+02   1.064 0.288321
factor(MOY)7    1.584e+03  5.824e+02   2.720 0.006879 **
factor(MOY)8    9.517e+02  6.372e+02   1.494 0.136240
factor(MOY)9    1.097e+03  7.399e+02   1.482 0.139214
factor(MOY)10   1.118e+03  7.282e+02   1.536 0.125564
factor(MOY)11   1.471e+03  8.040e+02   1.829 0.068292 .
factor(MOY)12   8.512e+02  6.861e+02   1.241 0.215615
factor(VISIT)2  1.531e+02  2.346e+02   0.652 0.514550
factor(VISIT)3  9.184e+02  2.665e+02   3.446 0.000643 ***
factor(VISIT)4  2.769e+03  4.798e+02   5.770 1.82e-08 ***
factor(VISIT)5  8.198e+03  1.143e+03   7.172 4.90e-12 ***
POP80          -1.950e-01  3.235e-02  -6.028 4.43e-09 ***
HHMEDI          5.769e-02  5.402e-02   1.068 0.286309
PERCAPI        -1.403e-01  1.578e-01  -0.889 0.374525
POVPR          -1.162e+02  4.916e+01  -2.363 0.018730 *
MFGPR           1.886e+01  1.361e+01   1.386 0.166782
COLLPR          6.577e+01  2.310e+01   2.848 0.004680 **
MAGE           -2.693e+01  4.735e+01  -0.569 0.569864
CART            4.787e-01  1.537e-01   3.113 0.002011 **
REAG           -6.215e-01  1.540e-01  -4.036 6.76e-05 ***
ANDR            4.724e+00  4.126e-01  11.449  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1815 on 330 degrees of freedom
  (40 observations deleted due to missingness)
Multiple R-squared:  0.8433,    Adjusted R-squared:  0.8314
F-statistic: 71.04 on 25 and 330 DF,  p-value: < 2.2e-16
```