Data Exploration

Upon initial exploration of the donor data, we observed relationships between the variables of interest and possible explanatory factors. Initially, we examined variables that might have the greatest influence on whether someone is a donor or not. Then, we applied similar techniques to identify the variables with the greatest influence on predicting the amount that identified donors can be expected to give. The data comprises a range of variables, including information about income and geographical location. The aim of the data exploration is to identify which variables have the greatest possible impact and to filter out noisy variables.
Upon initial inspection of household income (hinc), we observed that a significant portion of group 4 are donors, exceeding non-donors, with group 3 being a close second. This leads us to conclude that factors such as wealth are likely to impact donor participation. When examining the largest amount given and the most recent gift given, we observed a positive correlation, as expected, possibly indicating that the largest given amount could also be the first or most recent. However, the ideal target would be recurring donors, so the total number of donations would be an interesting addition.Following a similar pattern as wealth, homeowners are also much more likely to donate compared to renters. Another wealth metric shows similar results: when looking at low values, a higher low-income score has a lower likelihood of donating. Finally, when examining the number of children, we observed through a correlation plot that a higher number of children has a negative correlation with both donor participation and donation amount, making it the largest correlation outside of the wealth metrics. When examining donors and donation amounts, half of the dataset consists of non-donors, resulting in a large number of zero amounts, with an apparent normal distribution of amounts between $10 and $20. All figures can be found in the appendix of this report.
The data preparation was maintained via the given starter file. All values were standardized using the given process. NA values were dropped.

Methodology

For this assessment, We are looking to predict the amount of donation (DMAT) given that they are a donor. To do this, we are presenting three different models that aim to predict this amount, enabling decision-makers to identify and maximize the possible donation received by targeting these individuals. All of these models expand or build off of the given example model.

Model 1: Simplified model

The first model aims to simplify the example model by removing redundant variables identified in data exploration.

$$Yi = \beta_0 + B_i X_j$$

Xi is identification for each variate used; reg1, reg2 , reg3 , reg4 , home , chld , hinc , wrat avhv , incm , inca , plow , npro , tgif , lgif , rgif , tdon , tlag , agif.

Bi is the coefficient for each covariate.

In this simplified model, we aim to capture the region, home statistics, income statistics, and select giving statistics, giving a more simple interpretation of the model and its variates, without sacrificing too much of the models predictability.

Model 2: Feature Selection Model – leaps package

In Model 2, we aimed to take a different approach to selecting covariates using the leaps package and the regsubsets function. The variables from the dataset are systematically tested against the other possible models with the variables. This involves testing every possible model and comparing it to the possible models with different covariates. A limitation is that we are only able to use BIC. The following model is the result of this subsetting function.

$$Yi = \beta_0 + B_i X_j$$

Xi is identification for each variate used; reg1, reg2 , home , chld , wrat , incm, npro,tlag
Bi is the coefficient for each covariate.

Model 3: ZIP model – predicting both donor and amount

In the final model, we attempted to use a zero-inflated model to predict whether someone is a donor or not, and then the amount they donated if they were. This utilized the same variables as model 1. Questions did arise regarding the possibility of using a zero-inflated Poisson (ZIP) model, given its typical use when the response variables are counts. However, considering the large number of zeros in the non-donors and the range of donation amounts, as shown in the histogram of donation amounts (damt), we decided to attempt this approach.

Equation Format:

Combination of two components: a logistic regression for predicting excess zeros and a count model (Poisson).

$$\begin{aligned} \text{Count model:} \quad & \log(\lambda) = B_0, \\ \text{Zero-inflation model:} \quad & \log\left(\frac{p}{1-p}\right) = B_{0z}, \end{aligned}$$

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + B_i X$$

Xj is identification for each variate used; reg1, reg2 , reg3 , reg4 , home , chld , hinc , wrat avhv , incm , inca , plow , npro , tgif , lgif , rgif , tdon , tlag , agif.
Bi is the coefficient for each covariate.

Results

Evaluating the three models, we are using the metrics of mean squared error and mean absolute error. The best model was model 1 with an estimated profit value of 3720.44.

Model 1:

Model 1 performed the best out of our three models on the mentioned metrics. With a MSE of 1.89 and a MAE of 0.99, the model does a decent job at predicting the donor amount and ultimately was the best model in our investigation of the donor dataset. AIC is 6713. Adjusting for the over-sampling, the estimated profit value is 3720.44. The model can be seen below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.212210   0.047283 300.576  < 2e-16 ***
reg1        -0.066461   0.040173  -1.654  0.09821 .
reg2        -0.107764   0.043446  -2.480  0.01320 *
reg3         0.308424   0.041046   7.514 8.62e-14 ***
reg4         0.613848   0.042261  14.525  < 2e-16 ***
home         0.241538   0.061753   3.911 9.49e-05 ***
chld        -0.592173   0.038331 -15.449  < 2e-16 ***
hinc         0.495018   0.040499  12.223  < 2e-16 ***
wrat        -0.004035   0.042234  -0.096  0.92390
avhv         0.122656   0.041094   2.985  0.00287 **
plow         0.119287   0.045538   2.619  0.00887 **
npro         0.118582   0.045105   2.629  0.00863 **
tgif         0.088836   0.046719   1.902  0.05738 .
lgif        -0.053502   0.039102  -1.368  0.17139
rgif         0.508803   0.044632  11.400  < 2e-16 ***
tdon         0.077100   0.035528   2.170  0.03012 *
agif         0.671440   0.041176  16.307  < 2e-16 ***
```

Model 2: Feature Selection Model

Model 2 output metrics were disappointing with an MSE of 3.9 and a MAE of 1.5. When closing feature selection, we aimed to optimize the BIC metric which aims for a balance between complexity and explainability. Although it did not perform better than model 1, all but 2 of the variables were significant, highlighting that further fine-tuning and adding of select variables could drastically increase performance while maintaining explainability or a simpler model. When decreasing the number of variates, we also see an increase in the dispersion parameter to 3.19. AIC is 7990. Model output can be seen below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.909496   0.365122  38.096  < 2e-16 ***
reg1        -0.929269   0.110543  -8.406  < 2e-16 ***
reg2        -0.972559   0.097755  -9.949  < 2e-16 ***
home         0.869773   0.265245   3.279 0.001059 **
chld        -0.406247   0.037547 -10.820  < 2e-16 ***
wrat         0.003376   0.025060   0.135 0.892842
incm         0.005971   0.001496   3.991 6.81e-05 ***
npro         0.004538   0.001329   3.414 0.000654 ***
tlag         0.020084   0.013118   1.531 0.125936
```

Model 3: ZIP Model

Model 3 was the most adventurous, attempting to combine a count model with a continuous variable. The output metrics of this model showed it performing substantially worse, with a MSE of 172 and a MRE of 12.84. When looking at the summary output of this model, it appeared to be plagued by insignificant variables. The model output can be seen below.

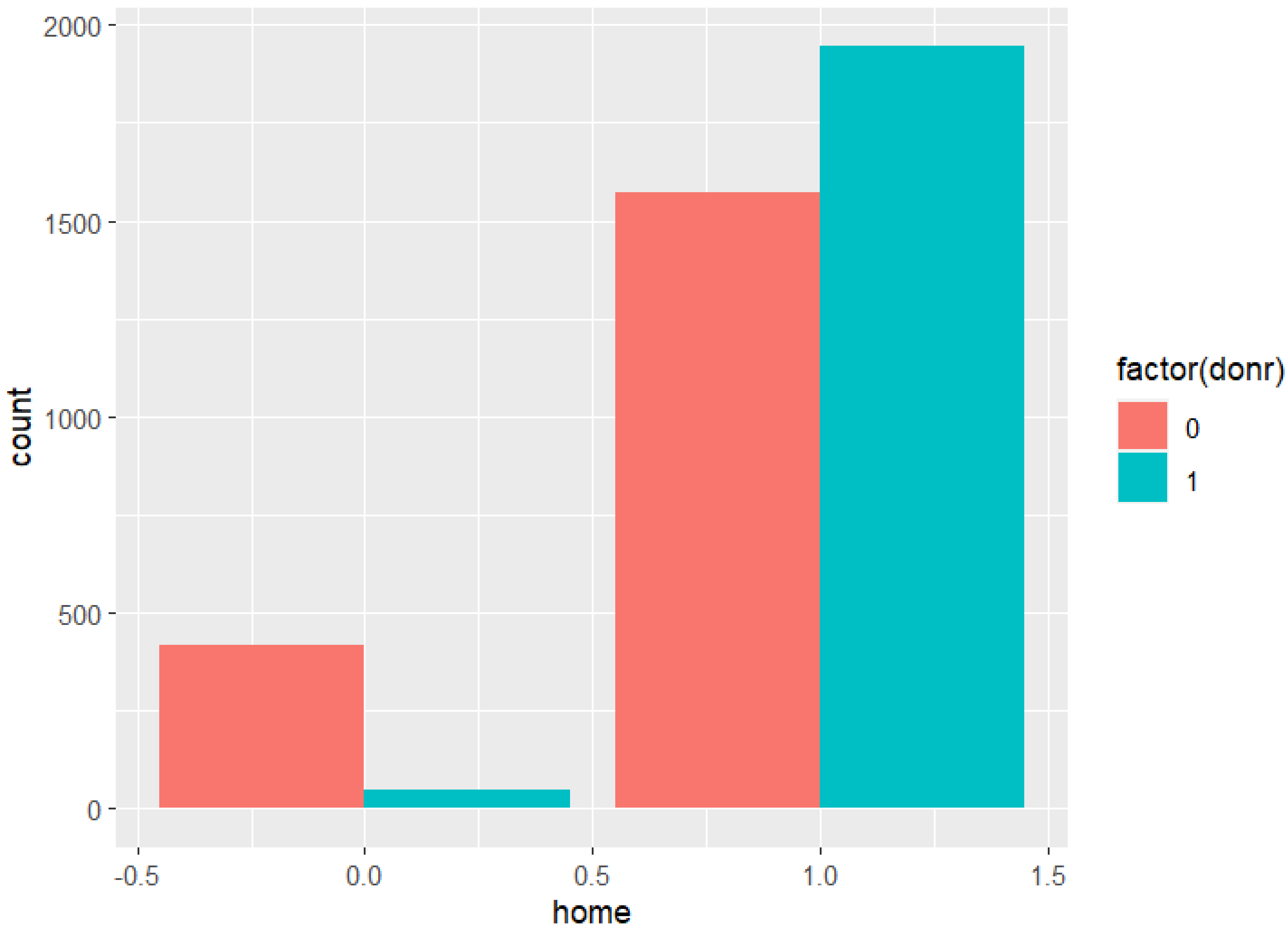
See appendix A.1…

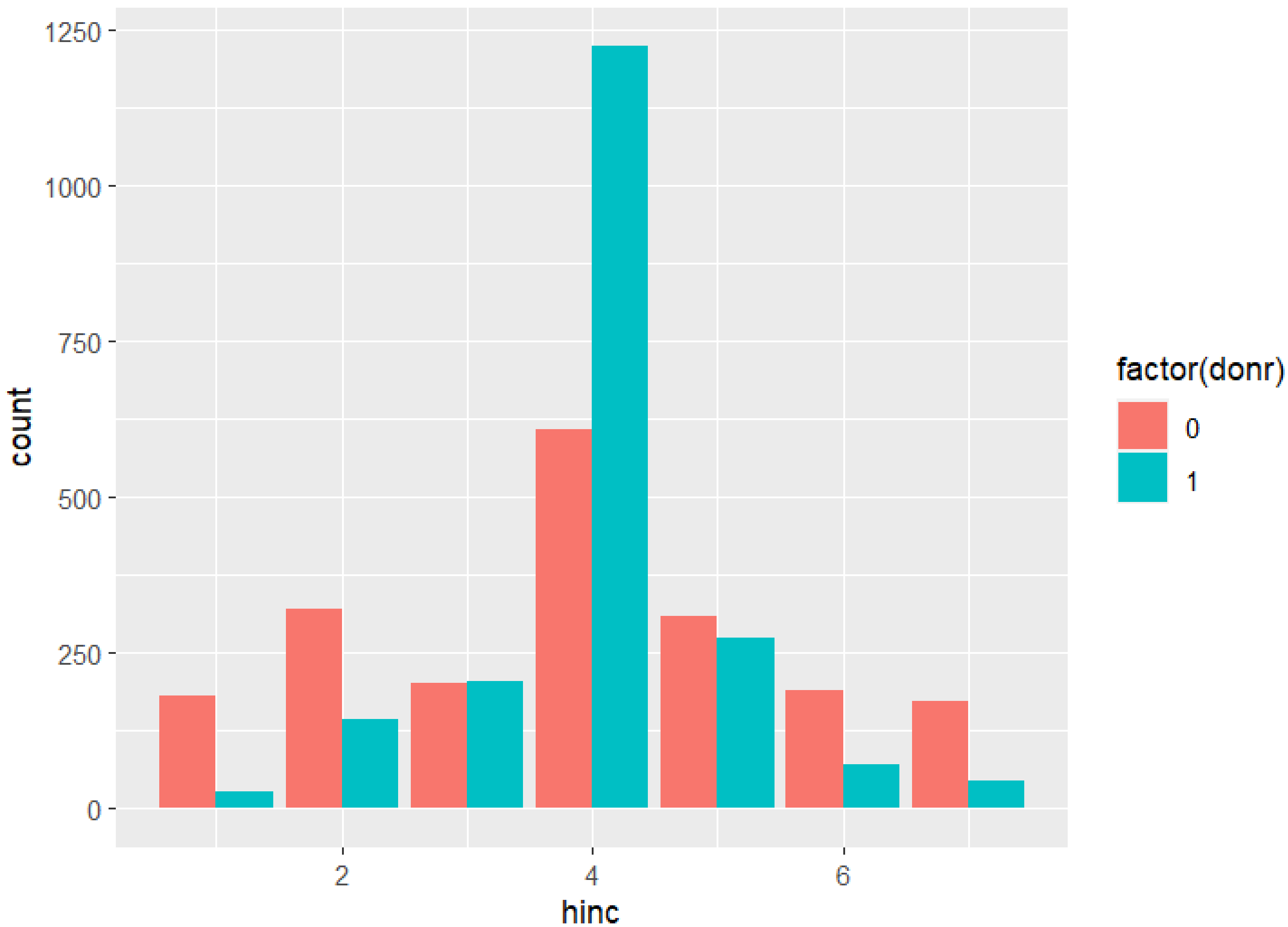
When looking at the histogram of damt, I wanted the ZIP to work, however, I am missing assumptions and the time to properly give the ZIP model a shot and see if it could work. Maybe there's no way it could. Not sure. Results were copied pasted, a grave sin, time crunch,

APPENDIX

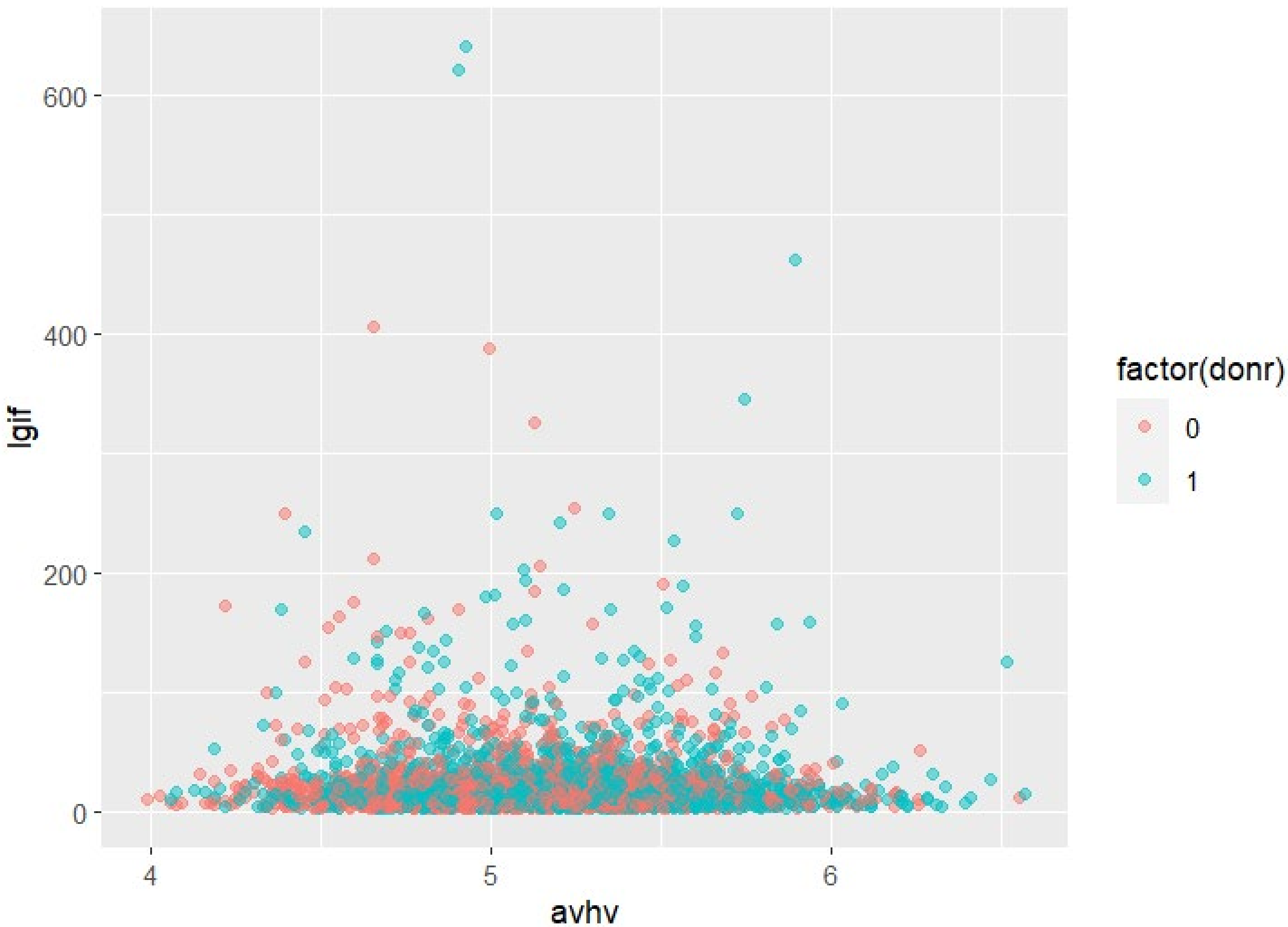
```
Count model coefficients (poisson with log link):
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.371e+00  1.602e-01  14.799  < 2e-16 ***
reg1        -7.156e-03  2.032e-02  -0.352 0.724678    
reg2        -1.157e-02  1.881e-02  -0.615 0.538421    
reg3         6.581e-02  2.484e-02   2.650 0.008056 ** 
reg4         1.202e-01  2.424e-02   4.958 7.11e-07 ***
home         5.255e-02  3.998e-02   1.314 0.188731    
chld        -3.034e-02  5.658e-03  -5.362 8.22e-08 ***
hinc         2.474e-02  5.874e-03   4.211 2.54e-05 ***
genf        -8.877e-03  1.202e-02  -0.738 0.460318    
wrat        -6.533e-05  3.680e-03  -0.018 0.985835    
avhv        -1.084e-02  2.978e-02  -0.364 0.715892    
incm         7.859e-04  4.839e-04   1.624 0.104307    
inca         1.212e-04  5.619e-04   0.216 0.829167    
plow         1.199e-03  7.432e-04   1.613 0.106761    
npro         3.267e-04  2.967e-04   1.101 0.270880    
tgif         4.273e-05  1.067e-04   0.400 0.688888    
lgif        -1.506e-04  2.329e-04  -0.647 0.517701    
rgif         2.649e-03  6.857e-04   3.863 0.000112 ***
tdon         8.110e-04  1.278e-03   0.634 0.525792    
tlag         3.333e-04  1.911e-03   0.174 0.861579    
agif         7.218e-03  1.221e-03   5.911 3.39e-09 ***

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  5.9283862  1.2645881   4.688 2.76e-06 ***
reg1        -1.3882154  0.1534160  -9.049  < 2e-16 ***
reg2        -2.5611416  0.1489810 -17.191  < 2e-16 ***
reg3        -0.0675973  0.1725975  -0.392 0.695319    
reg4         0.0379699  0.1705367   0.223 0.823808    
home        -3.4818610  0.2160876 -16.113  < 2e-16 ***
chld         1.3894692  0.0479528  28.976  < 2e-16 ***
hinc        -0.0599419  0.0371942  -1.612 0.107051    
genf         0.0707691  0.0989719   0.715 0.474583    
wrat        -0.3553164  0.0243921 -14.567  < 2e-16 ***
avhv        -0.2029870  0.2463941  -0.824 0.410036    
incm        -0.0150326  0.0042190  -3.563 0.000367 ***
inca         0.0002749  0.0049708   0.055 0.955893    
plow         0.0127967  0.0057199   2.237 0.025272 *  
npro        -0.0123733  0.0023107  -5.355 8.57e-08 ***
tgif        -0.0015921  0.0007855  -2.027 0.042667 *  
lgif         0.0009547  0.0020892   0.457 0.647703    
rgif        -0.0016976  0.0061871  -0.274 0.783797    
tdon         0.0413862  0.0090819   4.557 5.19e-06 ***
tlag         0.1279404  0.0143595   8.910  < 2e-16 ***
agif        -0.0047143  0.0104545  -0.451 0.652035    
---
```

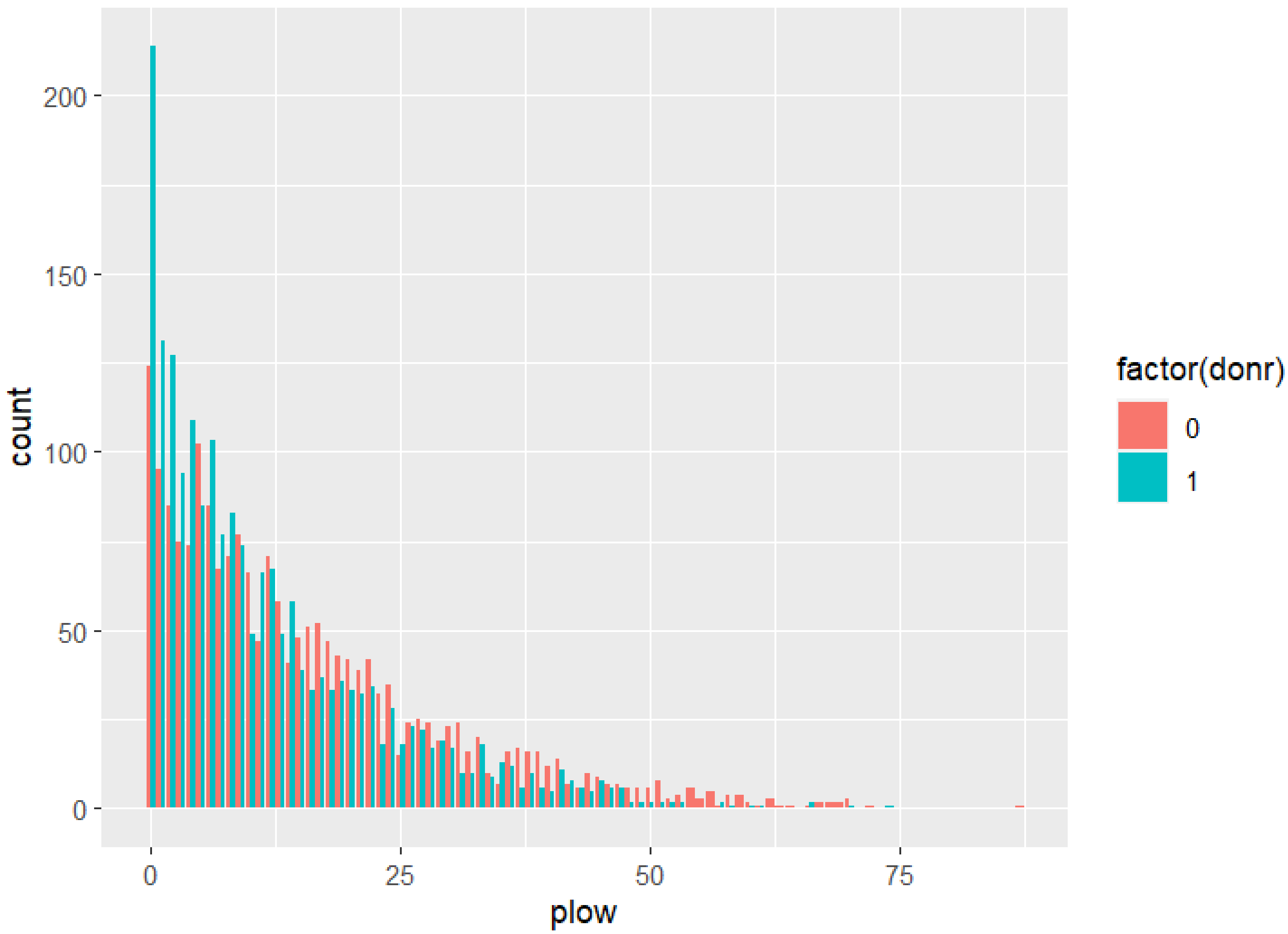

2000
1500
1000
500
0
count
-0.5
0.0
0.5
1.0
1.5
home
factor(donr)
0
1



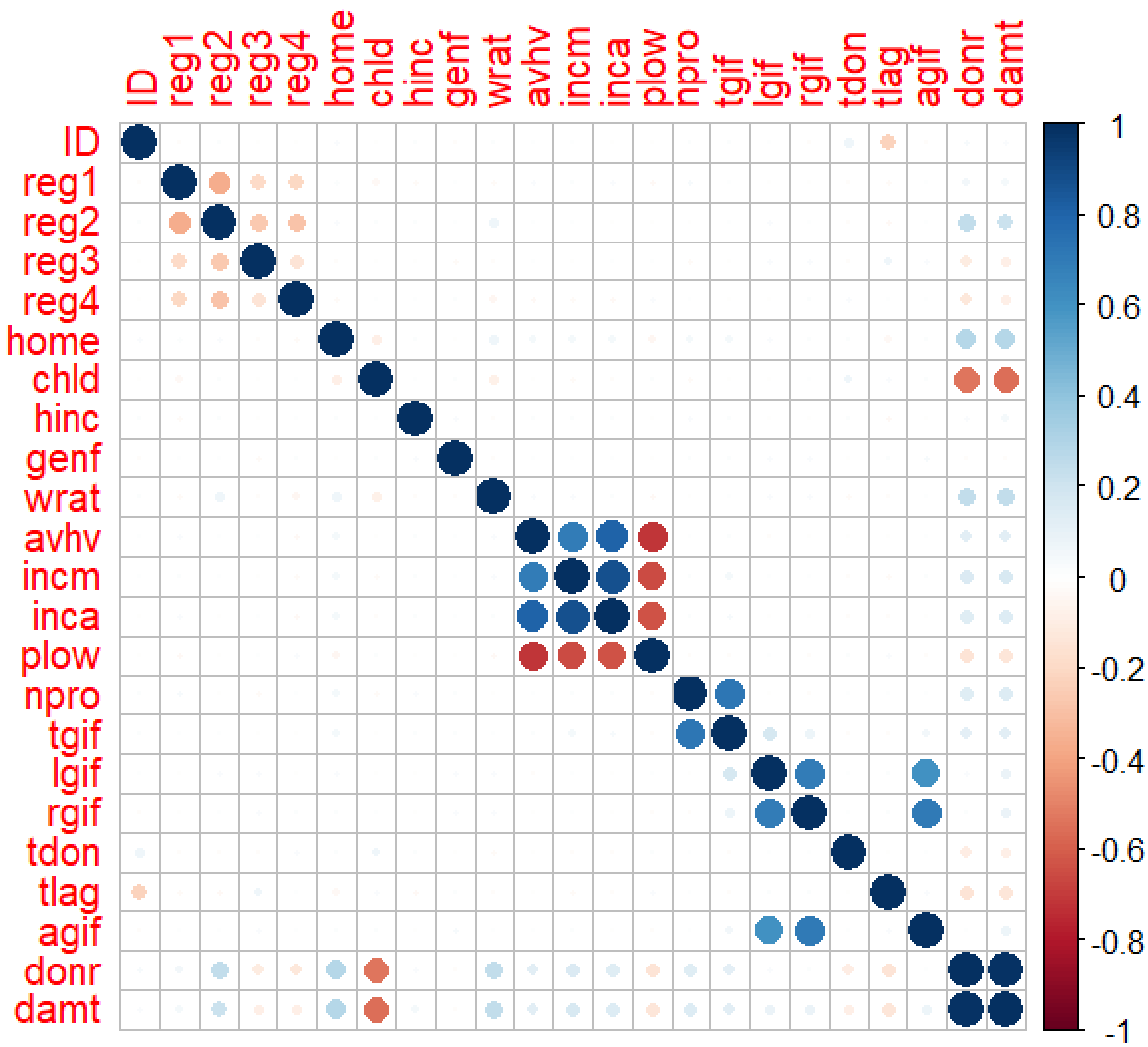
count
1250
1000
750
500
250
0
2
4
6
hinc
factor(donr)
0
1

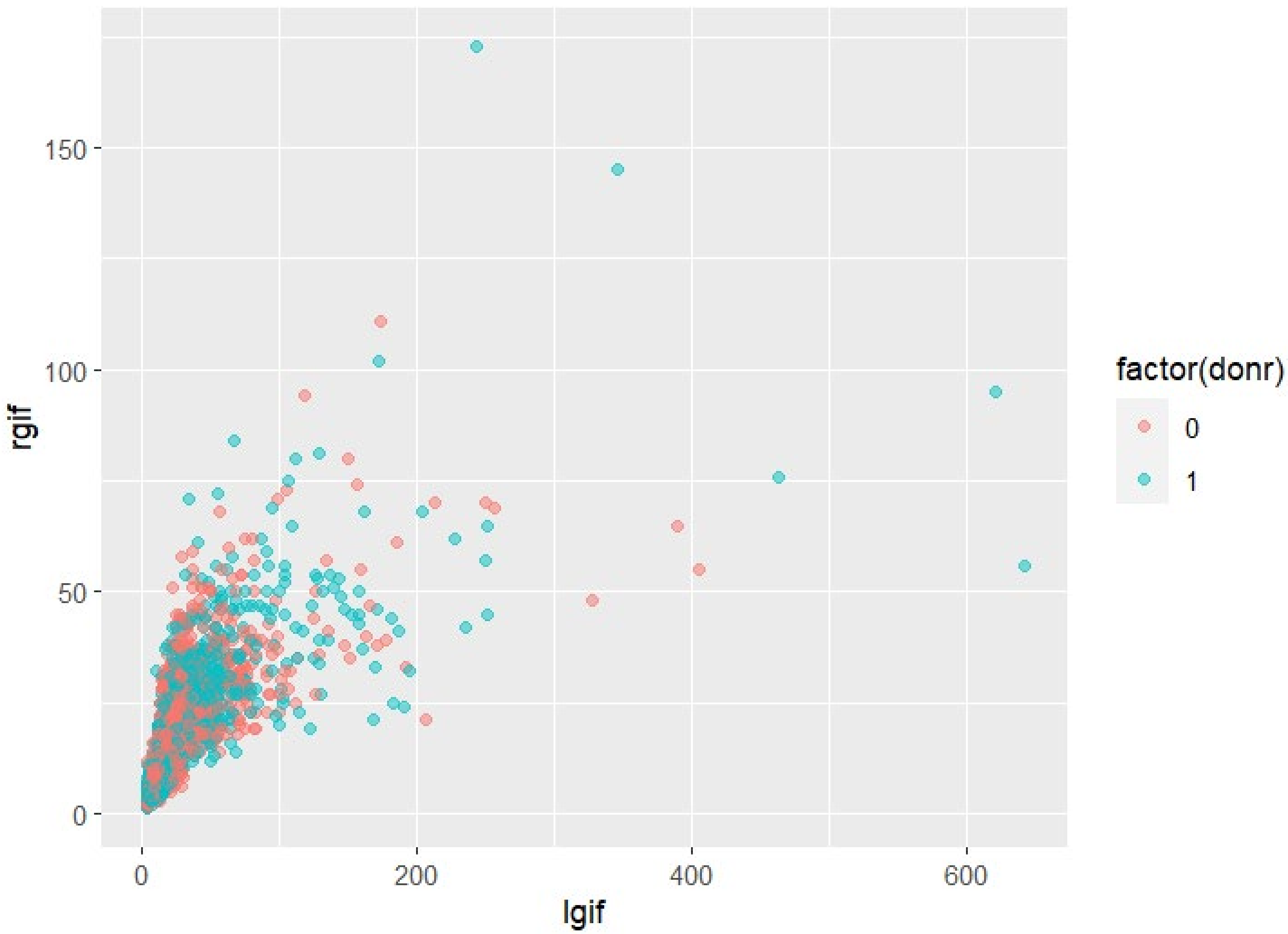


600
400
200
0
lgif
4
5
6
avhv
factor(donr)
0
1



200
150
100
50
0
count
0
25
50
75
plow
factor(donr)
0
1


ID
reg1
reg2
reg3
reg4
home
chld
hinc
genf
wrat
avhv
incm
inca
plow
npro
tgif
lgif
rgif
tdon
tlag
agif
donr
damt
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



150
100
50
0
0
200
400
600
rgif
lgif
factor(donr)
0
1