Charity Project

Author: Ricardo Herena

Introduction

Below we will explore our charity dataset with the mindset of solving a classification and a ev problem. According to our mailing data set, the average response is around 5.1% of those who responded the average donation was $15.62. Since cost to send a mail set costs $.99 it is implausible to send mailers to everyone.

As such we would like to develop models to identify good candidates for mailers, who are most likely to donate and donate large amounts. Our report is laid out below in order: data transformations / summary, General EDA, EDA for classification, EDA section for classification, and principle components analysis.

Data Transformations, Cleaning and Summary

After reading our data, we transformed DONR, HOME and HINC into factors. In addition we transformed all other text variables to factors. We also summarize the NA's per variable, data types, and the locational data including mean/median below.

```
##        ID              DONR              DAMT              AGE
##  Min.   :     1   Min.   :0.00000   Min.   :  0.0000   Min.   :20.00
##  1st Qu.: 47842   1st Qu.:0.00000   1st Qu.:  0.0000   1st Qu.:48.00
##  Median : 95799   Median :0.00000   Median :  0.0000   Median :62.00
##  Mean   : 95800   Mean   :0.05132   Mean   :  0.8046   Mean   :61.68
##  3rd Qu.:143718   3rd Qu.:0.00000   3rd Qu.:  0.0000   3rd Qu.:75.00
##  Max.   :191777   Max.   :1.00000   Max.   :200.0000   Max.   :98.00
##
##    HOME            HINC        GENDER          MEDAGE         MEDPPH
##  0   :15004   5      :13454   A   :    2   Min.   : 0.0   Min.   :  0.0
##  1   :46972   4      :10983   C   :    1   1st Qu.:38.0   1st Qu.:159.0
##  NA's: 8895   2      :10616   F   :38183   Median :41.0   Median :183.0
##               3      : 7189   J   :  290   Mean   :42.1   Mean   :186.7
##               1      : 7084   M   :30494   3rd Qu.:45.0   3rd Qu.:213.0
##               (Other):13427   U   :  741   Max.   :83.0   Max.   :650.0
##               NA's   : 8118   NA's: 1160
##     MEDHVAL         MEDINC          MEDEDUC         NUMPROM
##  Min.   :   0   Min.   :   0.0   Min.   :  0.0   Min.   :  6.00
##  1st Qu.: 524   1st Qu.: 237.0   1st Qu.:120.0   1st Qu.: 30.00
##  Median : 746   Median : 315.0   Median :120.0   Median : 49.00
##  Mean   :1063   Mean   : 346.7   Mean   :128.6   Mean   : 48.66
##  3rd Qu.:1205   3rd Qu.: 421.0   3rd Qu.:140.0   3rd Qu.: 65.00
##  Max.   :6000   Max.   :1500.0   Max.   :170.0   Max.   :194.00
##
##     NUMPRM12        RAMNTALL         NGIFTALL         MAXRAMNT
##  Min.   : 3.00   Min.   :  13.0   Min.   :  1.00   Min.   :   5.00
##  1st Qu.:11.00   1st Qu.:  44.0   1st Qu.:  4.00   1st Qu.:  14.00
##  Median :12.00   Median :  82.0   Median :  8.00   Median :  17.00
##  Mean   :12.99   Mean   : 107.8   Mean   :  9.91   Mean   :  19.88
##  3rd Qu.:13.00   3rd Qu.: 136.0   3rd Qu.: 14.00   3rd Qu.:  24.00
##  Max.   :64.00   Max.   :5674.9   Max.   :237.00   Max.   :1000.00
##
##     LASTGIFT            TDON           RFA_96
##  Min.   :   0.00   Min.   : 4.00   A1F    :17247
##  1st Qu.:  10.00   1st Qu.:16.00   A1G    : 7246
##  Median :  15.00   Median :18.00   A2F    : 4839
##  Mean   :  17.25   Mean   :18.42   A1E    : 4102
##  3rd Qu.:  20.00   3rd Qu.:21.00   F1F    : 3935
```

```
##  Max.   :1000.00   Max.   :27.00   A3E    : 2913
##                                    (Other):30589
```

```
##        ID      DONR      DAMT       AGE      HOME      HINC    GENDER
## "integer" "integer" "numeric" "integer"  "factor"  "factor"  "factor"
##    MEDAGE    MEDPPH   MEDHVAL    MEDINC   MEDEDUC   NUMPROM  NUMPRM12
## "integer" "integer" "integer" "integer" "integer" "integer" "integer"
##  RAMNTALL  NGIFTALL  MAXRAMNT  LASTGIFT      TDON    RFA_96
## "numeric" "integer" "numeric" "numeric" "numeric"  "factor"
```

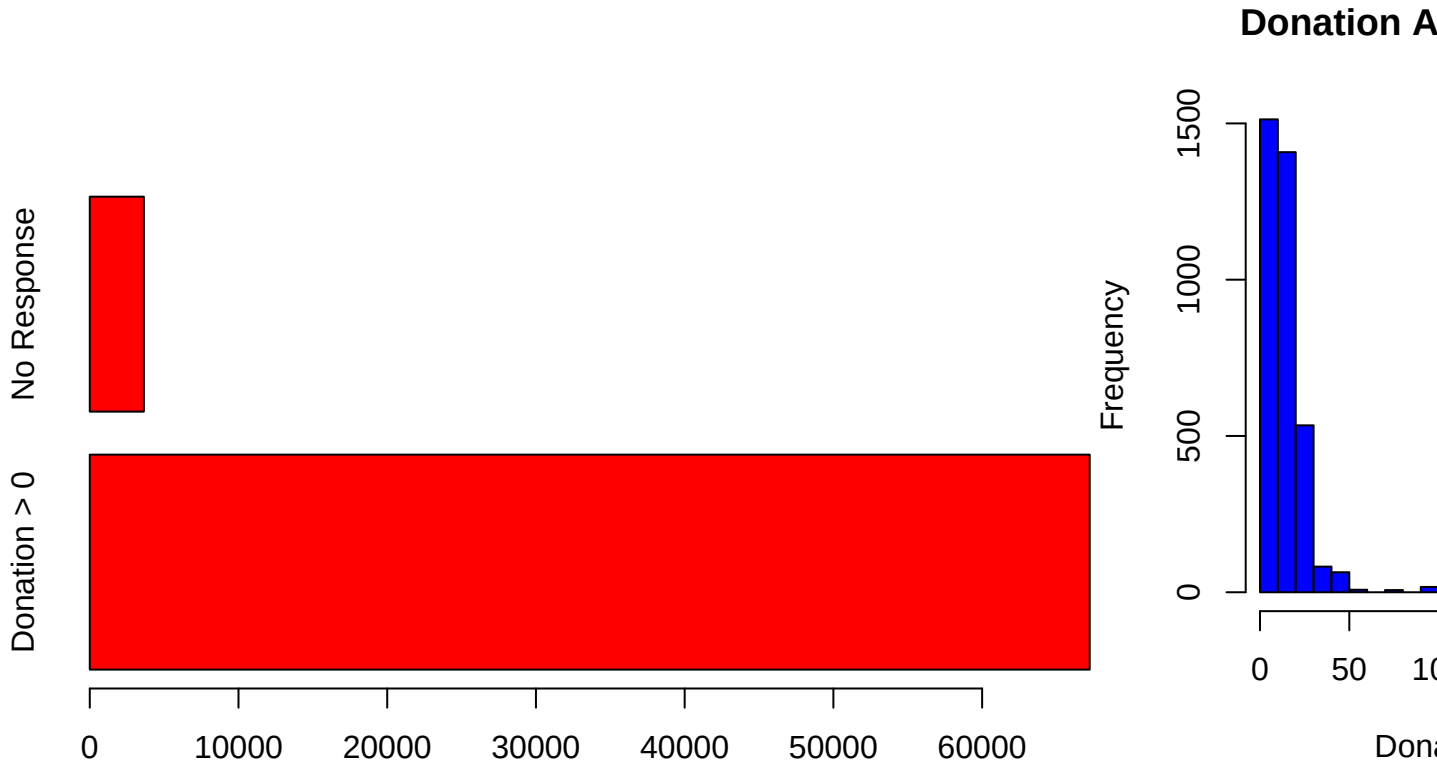
```
##          na_count
## ID              0
## DONR            0
## DAMT            0
## AGE             0
## HOME         8895
## HINC         8118
## GENDER       1160
## MEDAGE          0
## MEDPPH          0
## MEDHVAL         0
## MEDINC          0
## MEDEDUC         0
## NUMPROM         0
## NUMPRM12        0
## RAMNTALL        0
## NGIFTALL        0
## MAXRAMNT        0
## LASTGIFT        0
## TDON            0
## RFA_96          0
```

General EDA for Response Variables

Starting with our response variables DAMT we can see 67,234 of the DAMT are 0, and 3637 are greater than zero. For those that did respond the median DAMT was $14. Ranges from 1 - 200 with most values coming in 10 - 20.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.05132 0.00000 1.00000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   14.00   15.68   20.00  200.00
```



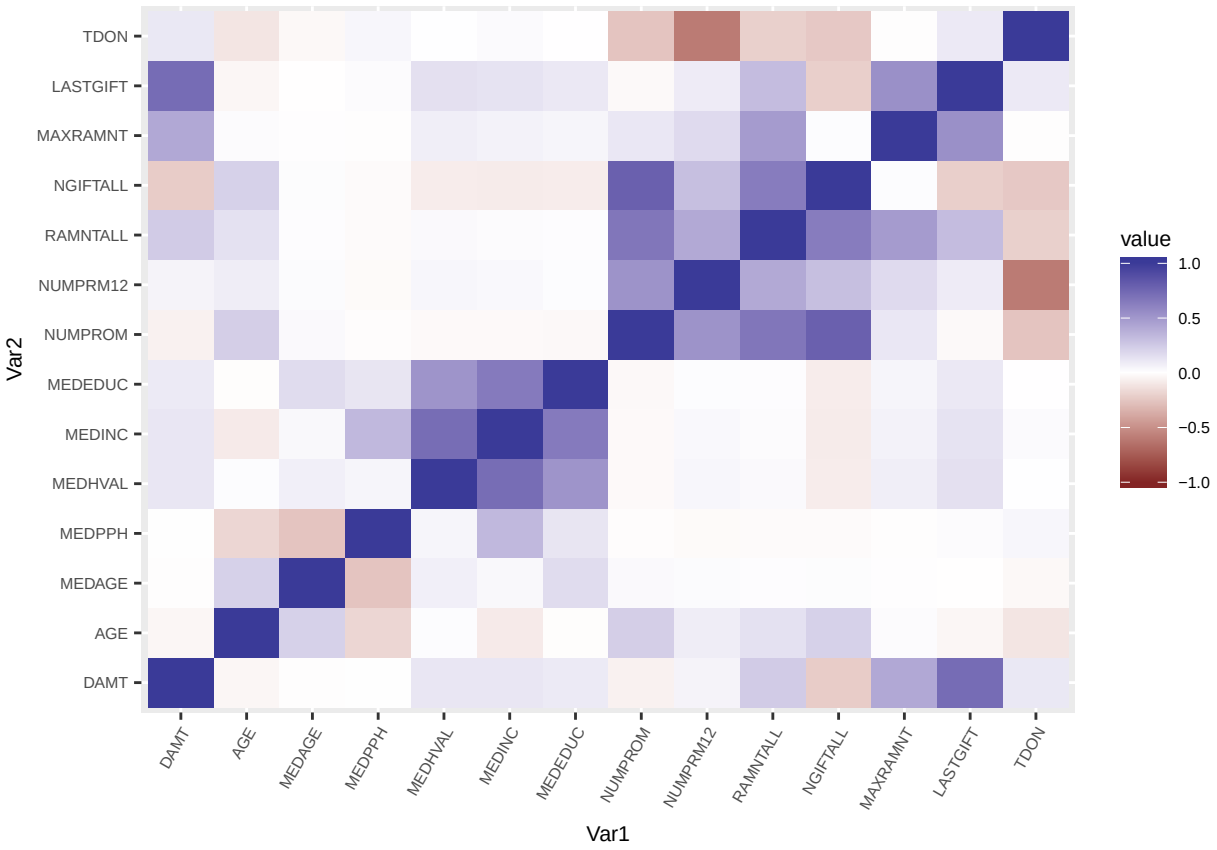
EDA for Regression modeling of DAMT

Below we have plotted the pairwise correlation of the numeric variables. This excludes, HOME, HINC, Gender and RFA_96.

We can see negative correlations to DAMT for: AGE, MEDAGE, NUMPROM, NGIFTALL.

We can see positive correlations to DAMT for: MEDPPH, MEDHVAL, MEDINC, MEDEDUC, NUMPRM12, RAMNTALL, MAXRAMNT, LASTGIFT, TDON.

```
##         DAMT          AGE       MEDAGE       MEDPPH      MEDHVAL
##  1.000000000 -0.038597430 -0.007067120  0.006517661  0.116305746
##       MEDINC      MEDEDUC      NUMPROM     NUMPRM12     RAMNTALL
##  0.116530060  0.098239525 -0.059724189  0.054758978  0.242751660
##     NGIFTALL     MAXRAMNT     LASTGIFT         TDON
## -0.226396326  0.412984842  0.722788561  0.108707853
```

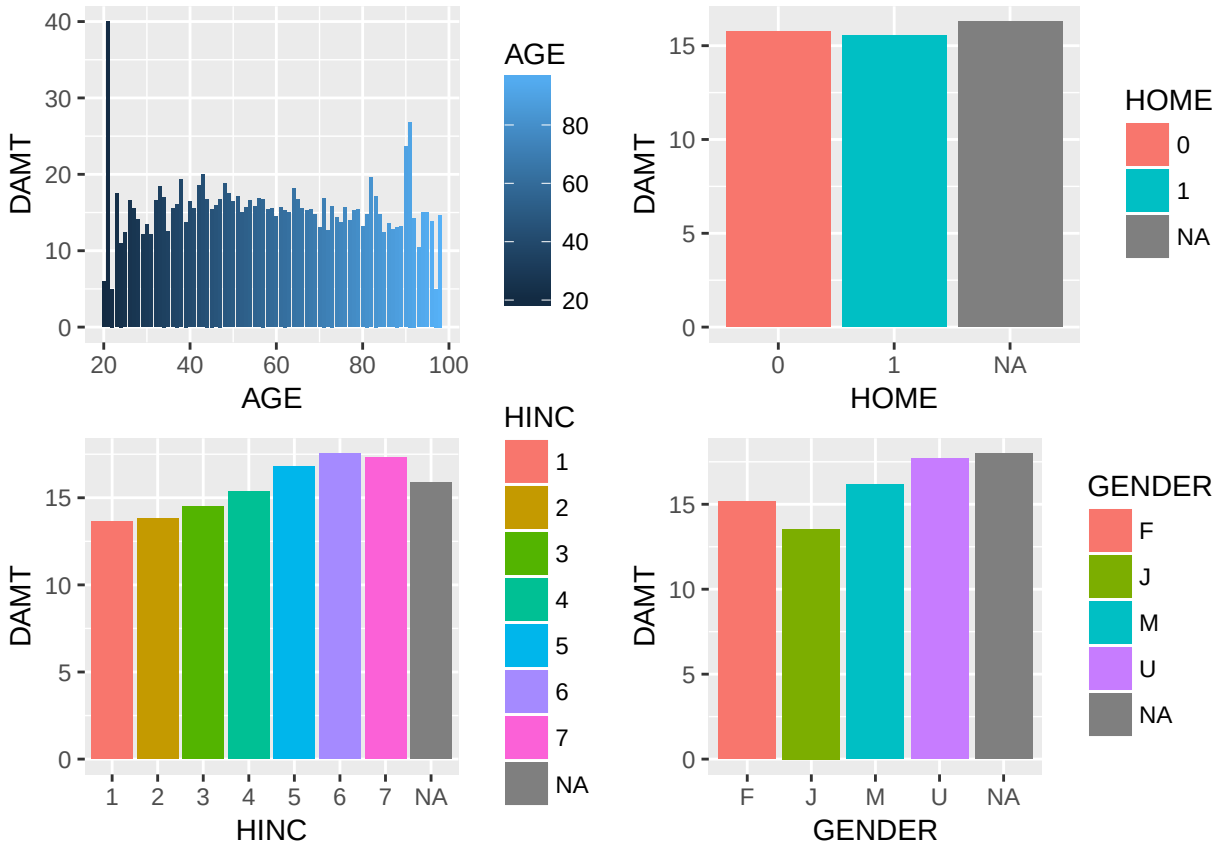


Next we will explore our basic donor characteristics including Age, Homeownership, Income bracket, Gender. For each we took the mean donation amount across the catagories listed.

Age and homeownership appear to have constant DAMT across categories.

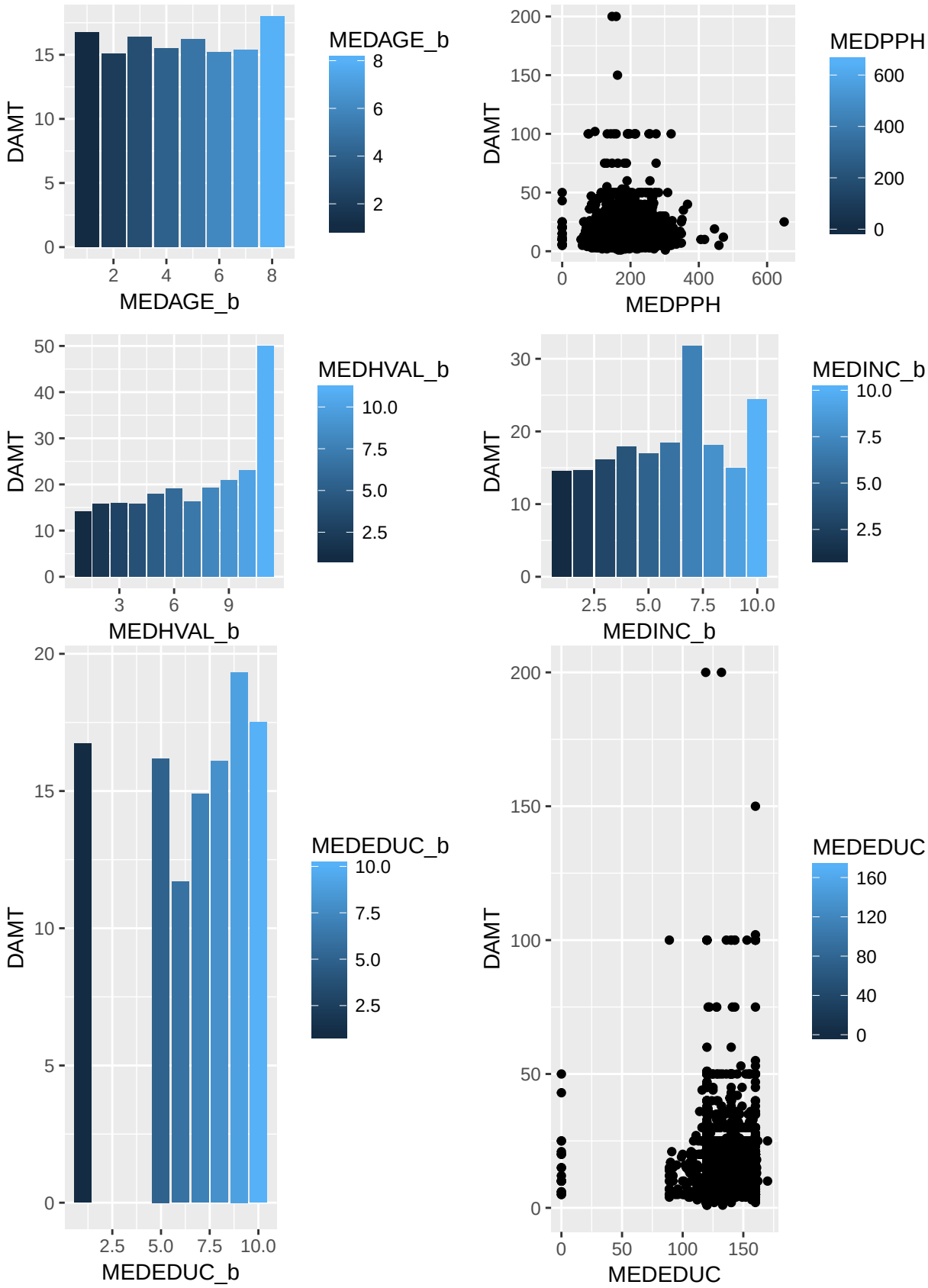
Gender appears to have some erroneous categories including A, but for M/F it appears to not make much of a difference. Importantly joint account holders are have a lower DAMT.

Income categories seem to have a positive relationship to donation amount.



Next we will explore our census data which reflects the donors nieghborhood characteristics. This includes median age, median person per household, median homevalue, median household income, median years of school completed. We bucketed many of these values into evenly sized buckets to try and produce means within each age, homevalue, income and years of school completed brackets.

Age, persons per household, income per household, appear to not have much of a relationship to donation amount. Where as homevalue and Homevalue appear to have positive correlations to donation amount.

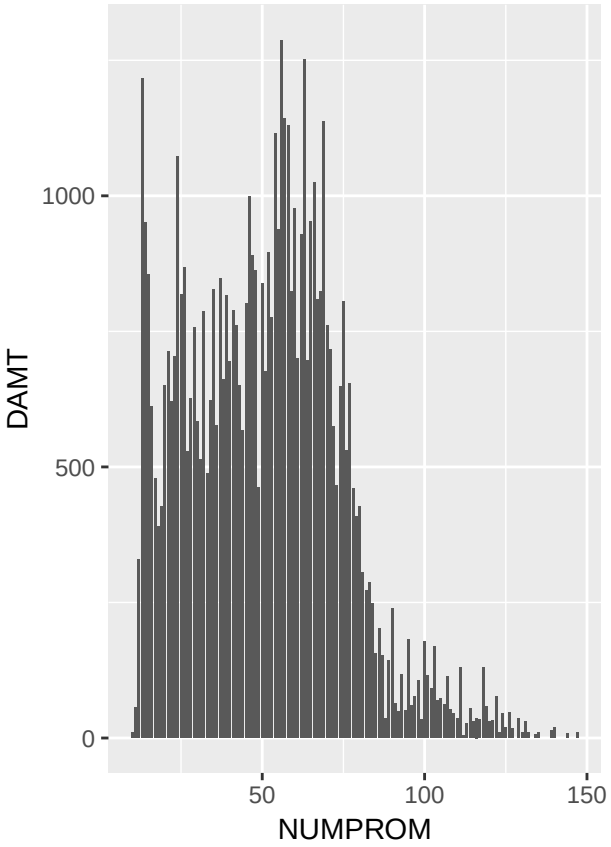


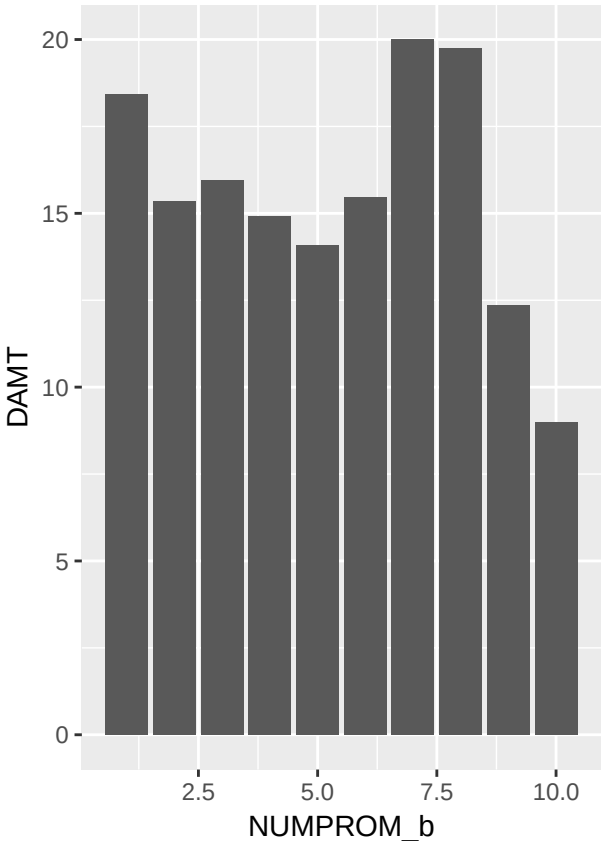
Finally we will explore promotion history file data. This includes lifetime number of promotions recieved,

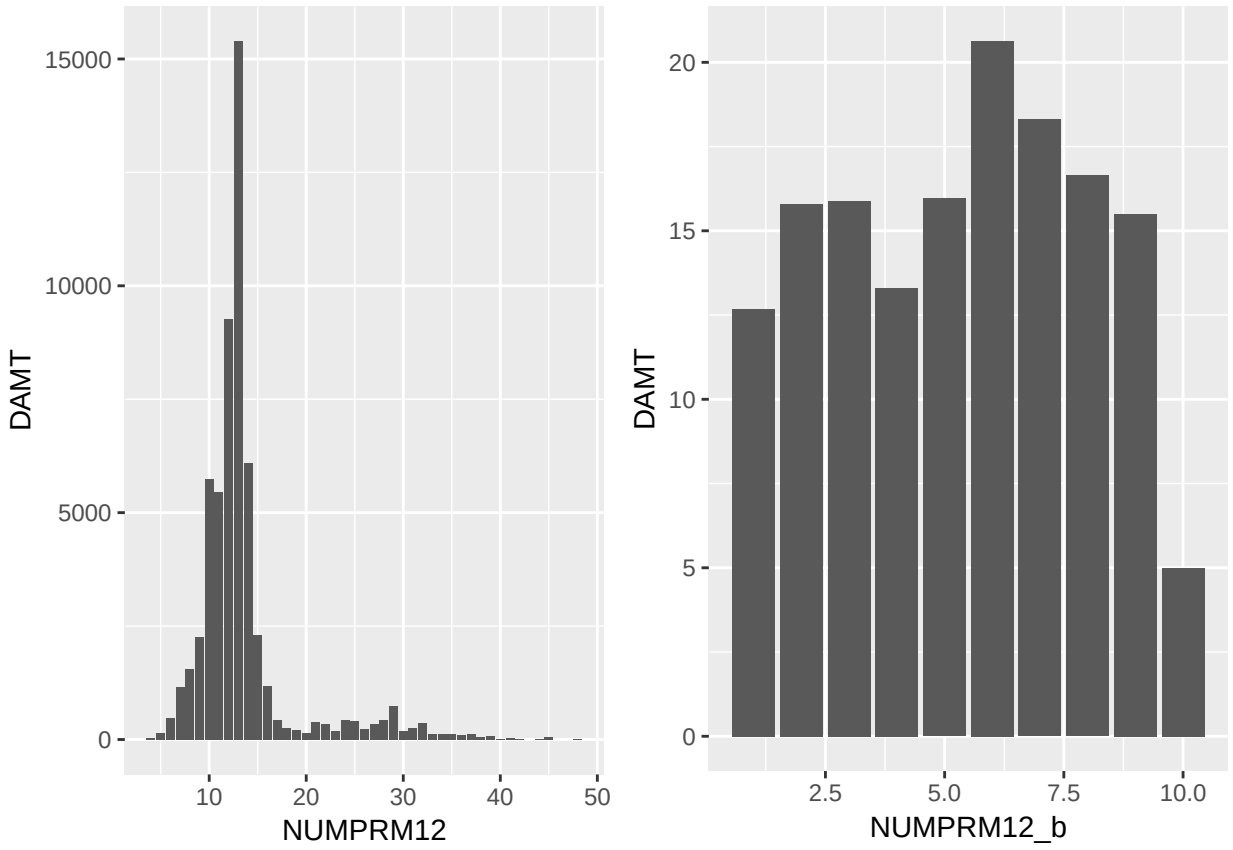
number of promotions recieved in the last month.

We bucketed both of these into logical order and present two plots for each. For the total number of promotions we noticed a negative trend, that seems to be broken in the middle with some sort of sweet spot, which equates to the 7th/8th bucket. This then falls sharply afterward indicating somesort of non-linear relationship.

For promotions in the last twelve months, it seems to increase intially then fall off rapidly. This may indicate the need to model these in two parts.

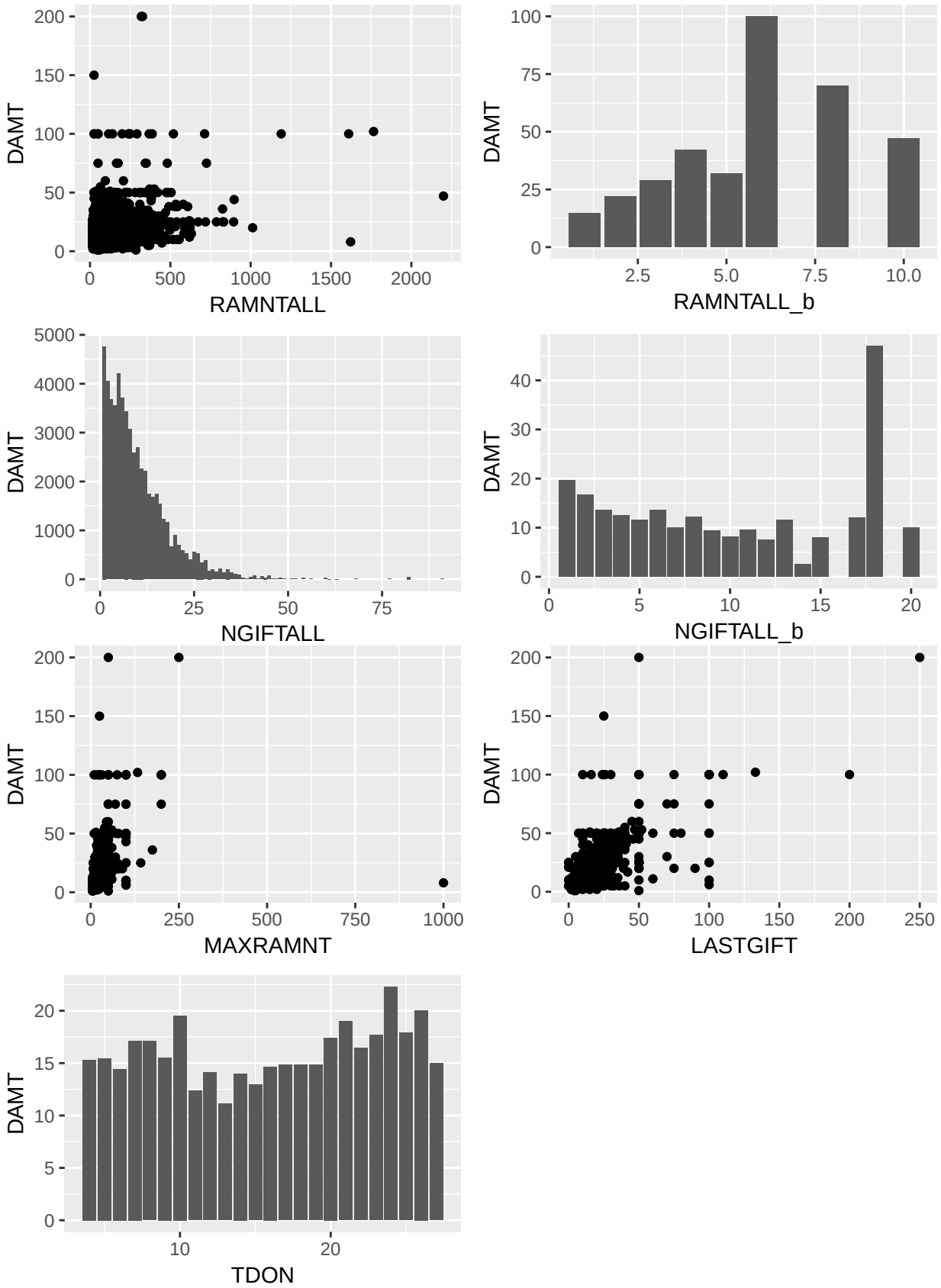






For this section we will explore giving history as it relates to donation amount. We strong positive relationship between individual donation amount and total lifetime donations. Conversely number of gifts is negatively related to donation amount. This may be that those with high lifetime donations make consistently above average donations, where as those who have made large number of gifts donate less on average.

Interestingly last donation amount and maximum donation both seem to positively related to donation amount. Number of months since the last donation doesn't seem to effect donation amount.

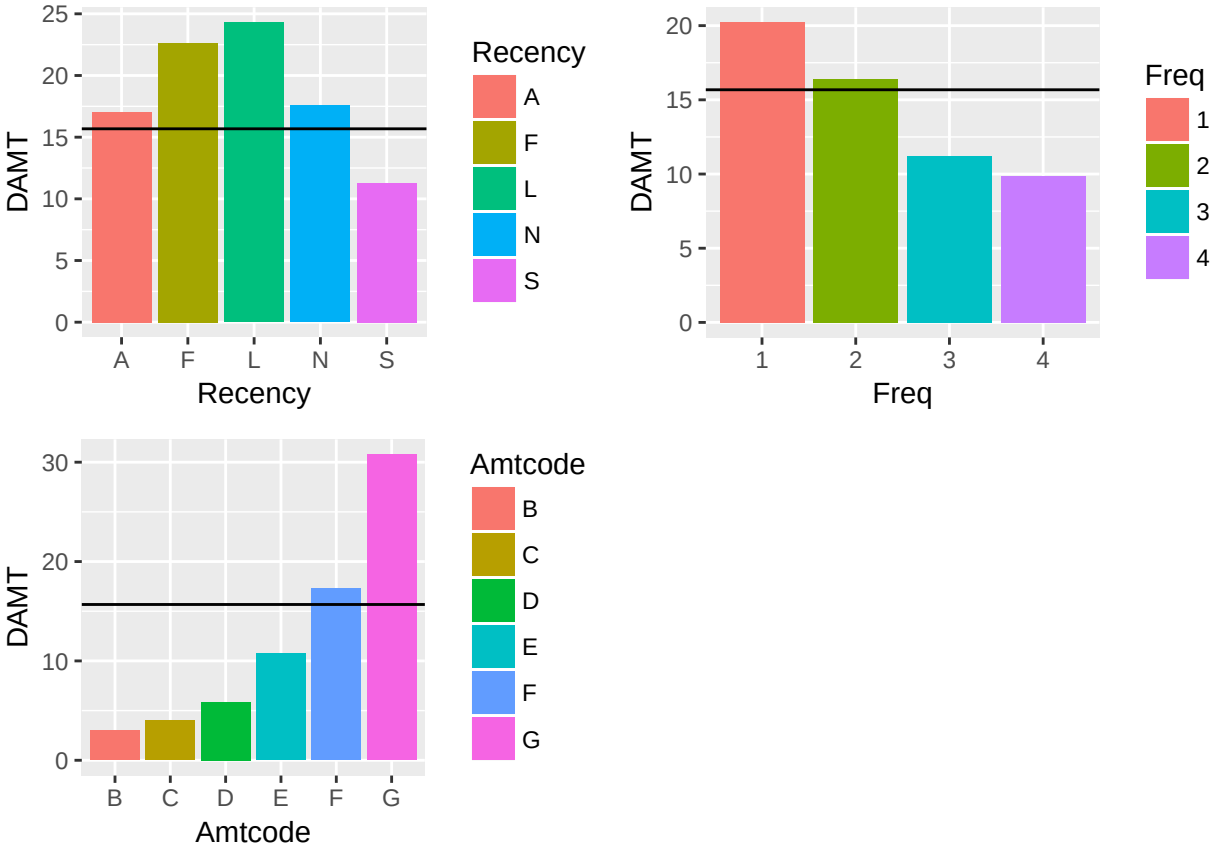


Lastly we will explore recency, frequency and amount of donations and how they effect donation amount.

The different codes correspond with different types of donor frequency.

As we can see recency indicates active ("A") donors, New ("N"), First time donors("F") and ("L") Lapsing donors donate above average amounts. Where as ("S") or star donors seem to donate less on average. This is consistent with two groups of donors, those who donate above average less often, and those who donate lower amounts more often.

Frequency seems to tell the same story with a steep negative correlation to donation amount. Where as amount of the last gift seems to have a strong relationship with donation amount. Those who settle on an amount range are likely to stick to it.



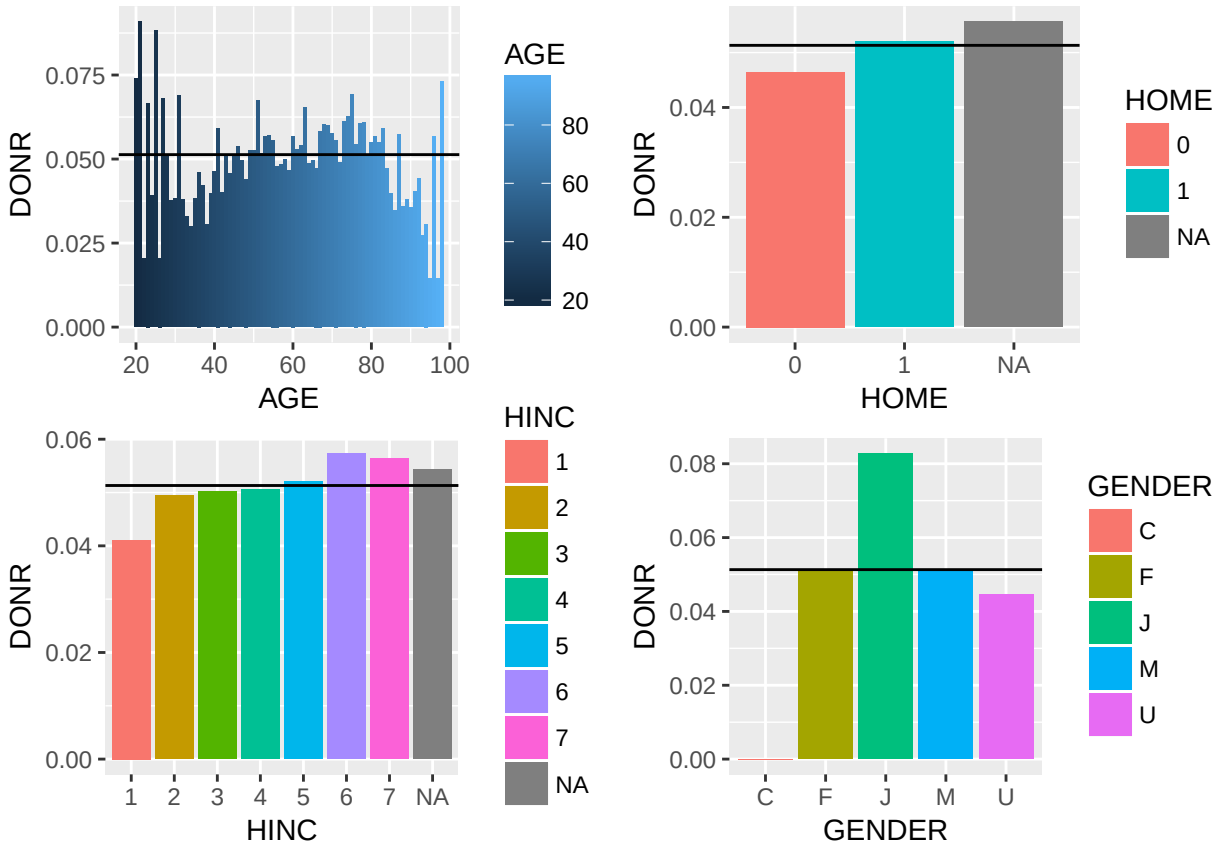
EDA for Classification modeling of Donation Probability

Next we will explore the same or similar variables on the whole dataset with the goal of identifying who is most likely to donate, IE has the highest probability of DONR = 1.

As we can see age seems to follow a non linear relationship, with less donations from those 20 - 40, and more than average response through 40 - 80, and then tailing off again.

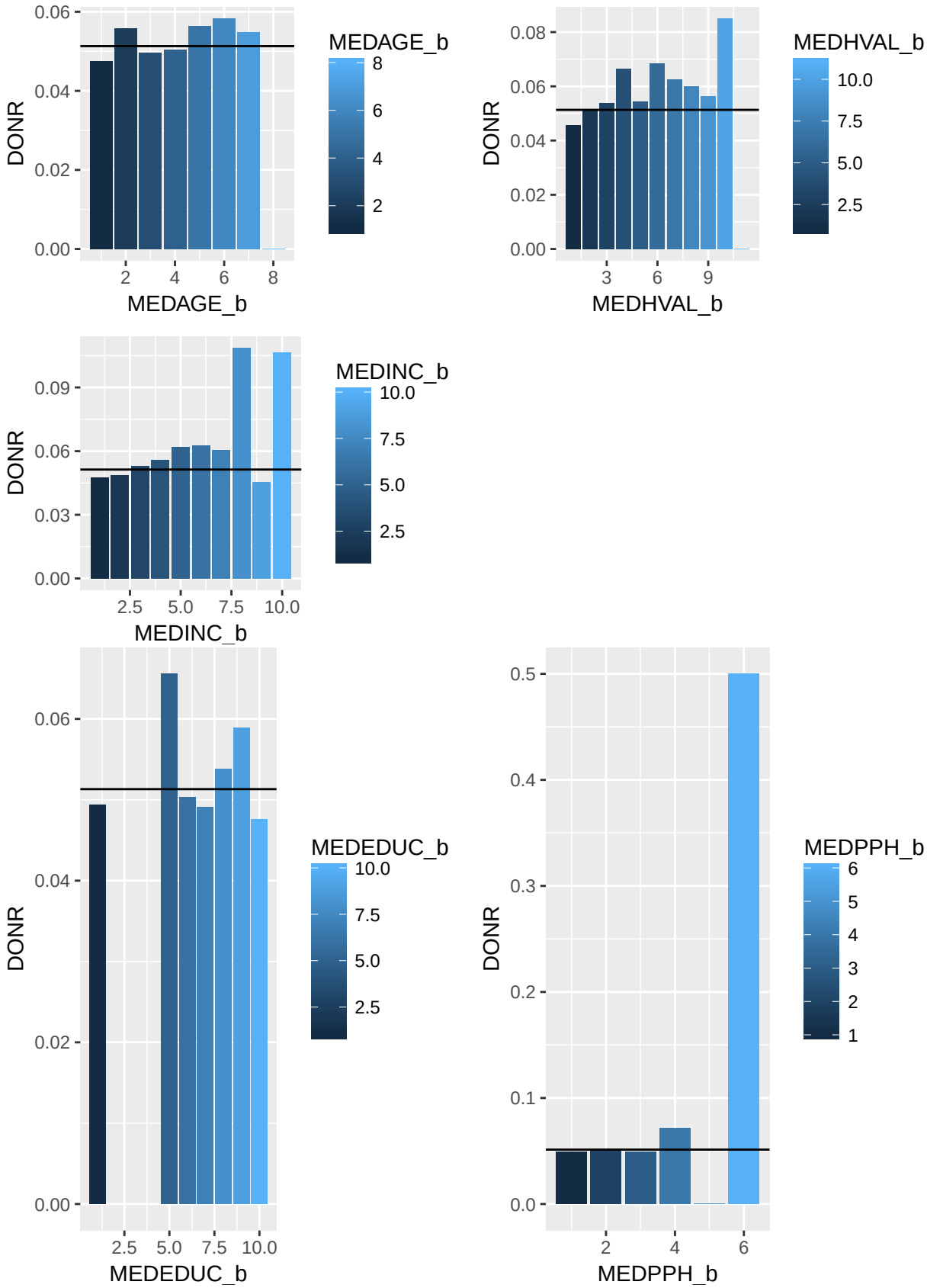
Household income was positively associated with response rates.

Homeownership, and gender didn't seem to indicate higher reponse rates.



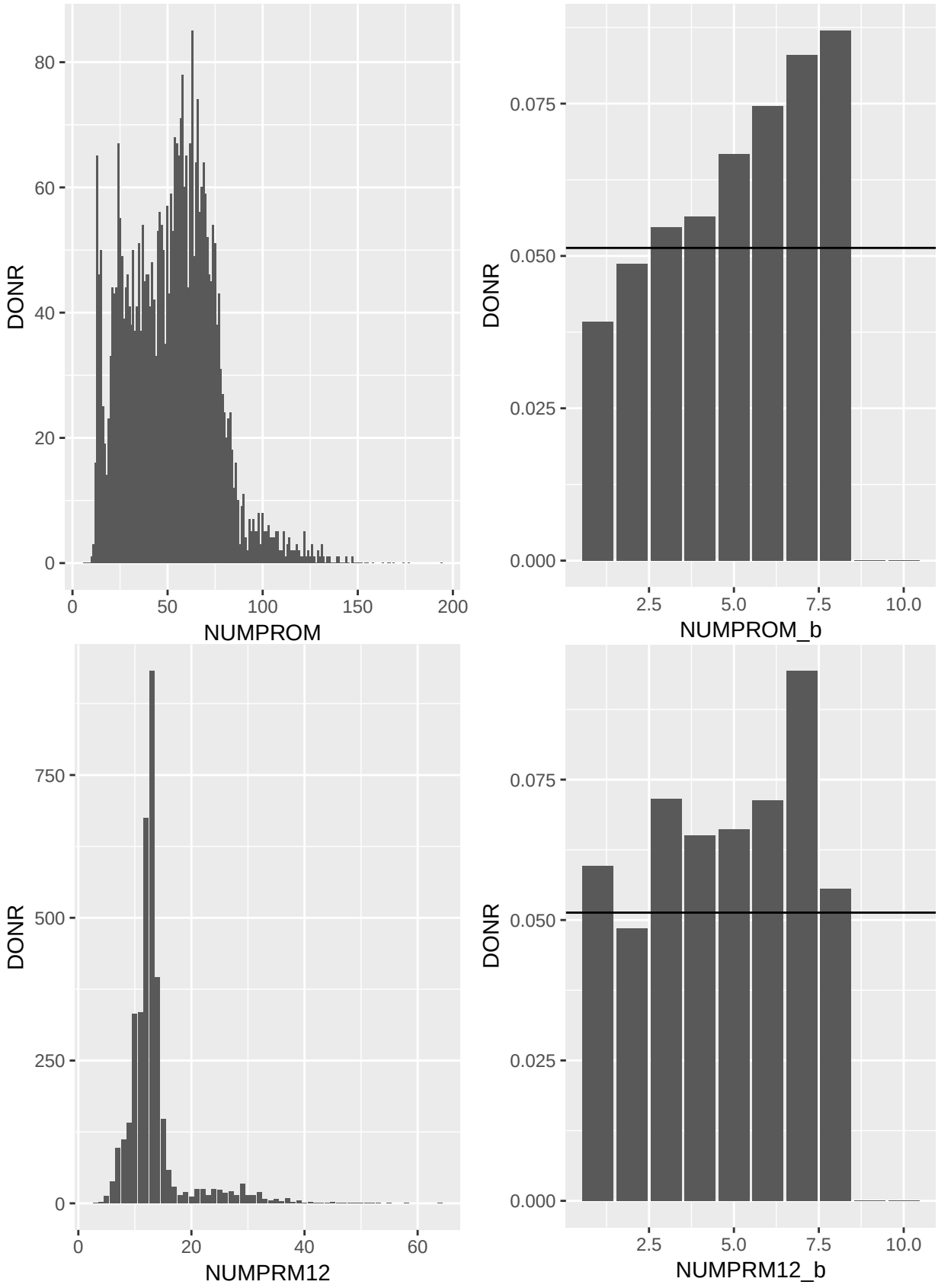
Next we will explore neighborhood characteristics for response rates. We can see age, median income, and home value seems to positvely associated with response rates.

Education and persons in the household seem to be a mixed bag with no descernable pattern.



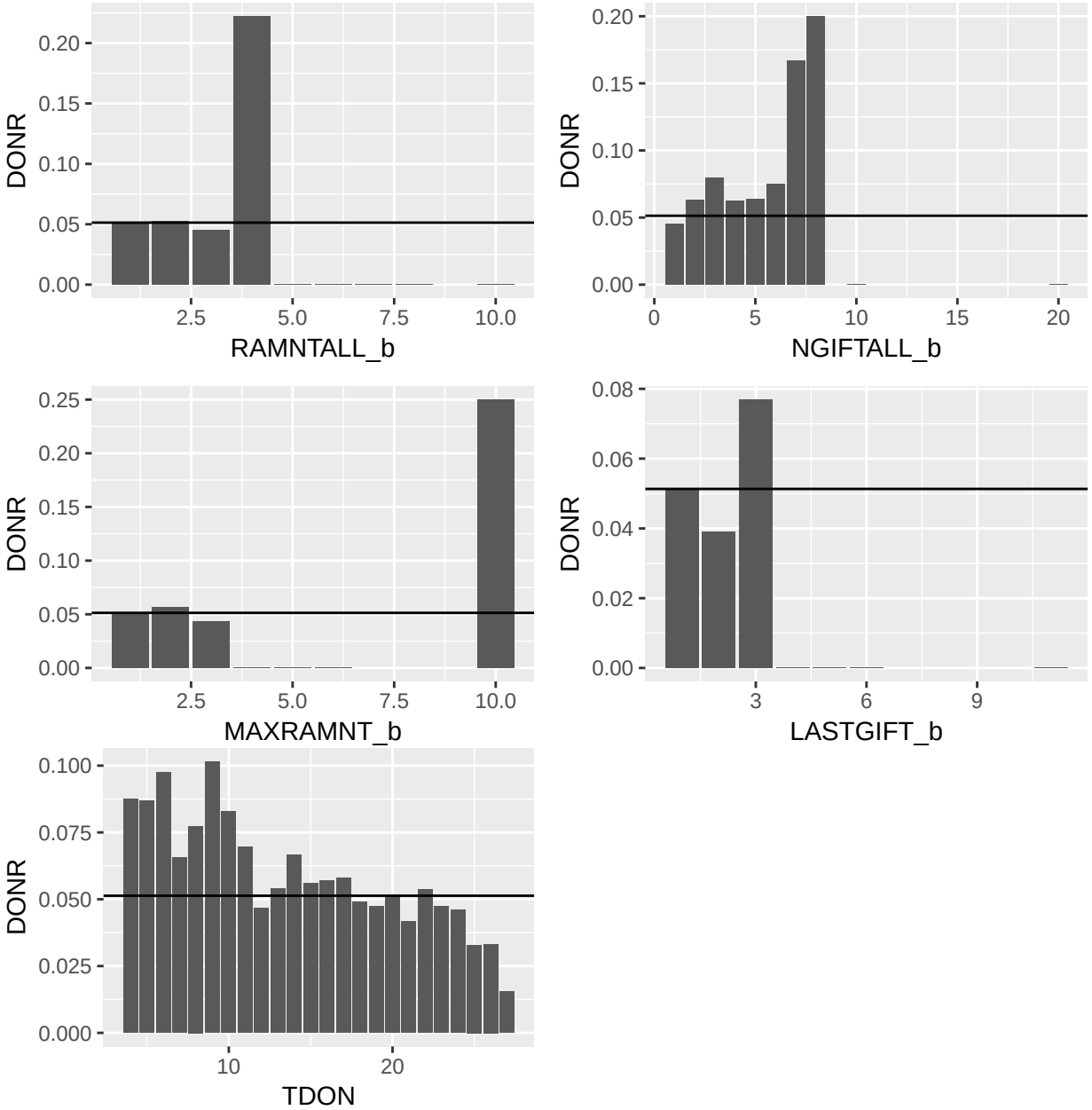
As we can see from the promotions history data, those who have been targeted before are more likely to

respond to a future promotion.

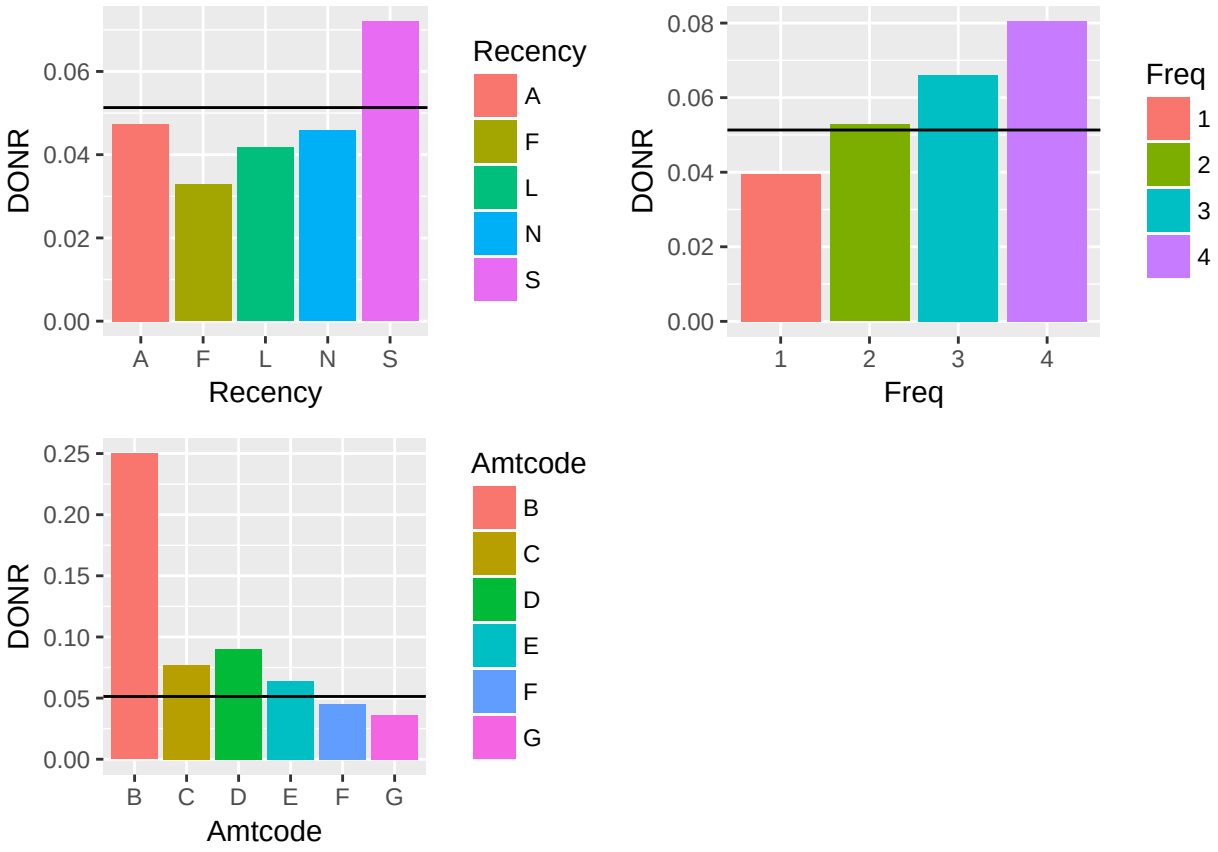


When we look to our giving data and response rates we see a mixed story. Donation amounts, lifetime donations, and last donation amount don't seem to have much relationship to response.

Total number of gifts is (NGIFTALL) positively associated with response rates, where as number of months sinc the last donation has a steep negative relationship to response rates.



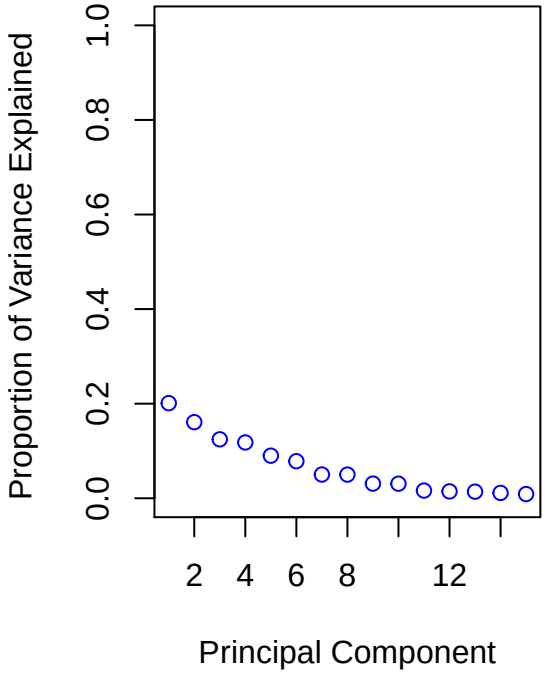
As we can see when it comes to frequency, those who have responded more often are more likely to respond, interestingly only star donors were above average responders. Finally, our small donation responders are by far more likely to respond to future campaigns.

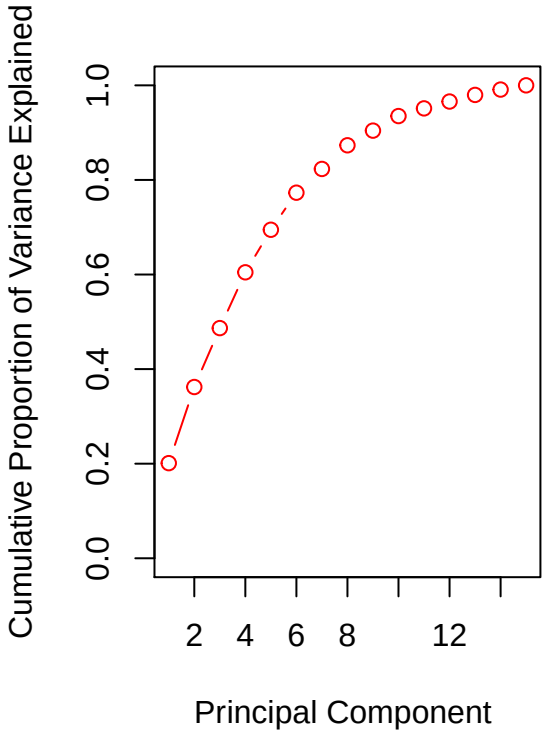


Principle Components Analysis

Below we will conduct PCA on our initial datasets. Please note the only variables included in this dataset are the quantitative variables. As we can see our scree plots seem to indicate incremental gain from most of the principle components, we don't see a sharp drop in proportion of variance explained. It is interesting to note it takes 8-9 principle components to explain the variation in the data.

This may indicate that there is information data from most of the variables, and that we will need a large combination to explain both classification and regression.


Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
12
Principal Component


Cumulative Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
12
Principal Component