**Executive Summary**

**Project Goal**

The goal of this project is to predict the probability that a match will occur between two people who are customers of a speed dating service. This knowledge can be used to increase the number of matches made per session. This should help to raise the customers initial satisfaction and hopefully encourage them to continue to pay to use the service.

**Data**

8378 customers responded to 195 questions about themselves and the speed dating service. This was winnowed down through several iterations into a data set comprised of 7 variables.

* matchYN: is the dependent variable that will be predicted. It is a factor, either “yes” or “no.”
* intRate: correlation between participant’s and partner’s ratings of interests. It is a number.
* agePart: the age of the person you are meeting. It is an integer.
* age: the respondent’s age. It is an integer.
* field: area of respondent’s undergraduate study. It is a factor with 18 levels.
* almaMater: institution attended for undergraduate degree. It is a factor with 11 levels.
* dateFreq: how often the respondent goes out on dates. It is a factor with 7 levels.

To see the full list of variables see appendix 1.

**Approaches**

The first step was to set the seed to make the results replicable. The next step was to partition the data. Since this was a large dataset 80% of it was used to train the models and 20% was reserved for a test set. Then a variety of different models were examined. They attained high accuracy but did not predict the goal, matchYN very successfully. In order to accomplish the goal oversampling was employed.

**Results**

The first logistical regression model correctly predicted 2/270 yes matches on the test data. After oversampling it was able to correctly predict about 55% of the yeses on test data. Random forest models achieved similar results ( 54% of yeses predicted on test data). Overall accuracy was lowered, but the stated goal of predicting yeses was somewhat successful.

**Discussion**

I was unable to both cross validate and generate a confusion matrix with the logistical regression model. It was a large dataset however and the results were similar to the cross validated random forest model. There are almost certainly good variables remaining in this set to train the models on that were discarded. I would also like to examine some further modeling methods such as SMOTE, or other forms of machine learning. It would be interesting to ensemble the models together to further refine the ability to predict yeses. This method would allow the strengths of the various models to be exaggerated while their weaknesses could be diminished.

**Report**

**Tidying**

The first and most labor-intensive step of the project was to tidy the data. With such a large data set it was not feasible in the time allotted to examine each variable individually, or to model on the entire set. I selected a refined set of variables to begin with. They were gender, matchYN, intRate, agePart, age, field, almaMater, brains, home, goal, dateFreq, socialScore, attitude, and swagLevel. I took a closer look at these variables and determined swagLevel had too many NA’s, so I removed it. This was the list of variables I started with.

Now some of the variables had to be regrouped. almaMater, for instance, was a factor with 245 levels. I cleaned that up by grouping them into 11 categories, "Other," "Asian School" for schools in Asia, "BigPrivate" for large private schools, "CaliSchool" for schools in California, "European School" for schools in Europe and Russia, “Foreign American School" for North and South American schools outside the USA, "Ivy" for Ivy League schools, "Ladies" for Female only/formerly so schools, "LittleIvy" for the Little Ivy League schools, “PublicSchool" for the public schools, and “SmallPrivate" for the small private schools.

I then grouped the different scores for intelligence into 3 levels, “low,” med,” and “high.” Subsequent evaluation showed a lack of variance. The values were based on SAT scores, so I tried a larger number of levels and grouped them by 100’s (i.e. 1000-1100, 1101-1200, etc.). This also proved to be ineffective. This variable was revealed to be noise and was later discarded.

I then correctly identified the variables as factors, numbers, and integers. I also fixed all the labels on the levels of the factor variables. At this point I wanted to look closer at the variables. I created graphs that displayed the percentage of matchYN responses for the various levels of the factor variables. I then created a logistic regression model and did backwards selection. I used understanding gained from the graphs and the AIC information from the backwards selection process to trim the data further until I reached my final dataset. The final variables were matchYN (dependent variable), intRate, agePart, age, field, almaMater, and dateFreq. Most of the last cut were simply noise. The variable social was eliminated because it was strongly correlated with dateFreq (they were both measures of how often you go out) and DateFreq had higher variance. The tidying was finally complete.

**Modeling**

The first step in training the models is setting the seed to make your results replicable. Then I split the data into a training set and a test set. It was a large database, so I went with an 80% training 20% testing data split. At this point I also loaded in some functions to calculate RMSE and MAE. I then trained a logistic regression model on the training data. The good news is that the model was very accurate. This is because only about 16% of the responses were yes. If the model just guessed no every time then it would be 84% accurate, for instance.

The solution to this issue is to oversample the data. Using a package called ROSE I oversampled, under sampled, and did a combination of both to the data to raise the proportion of the yes answers to 50% in the training data. It is important to note that I did this after I had partitioned the data. If I had not, then some of the yeses from the test set could have been included in the training data and influenced the modeling.

I trained a logistic regression model on the three different methods of resampling. It turned out that the oversampling method was slightly more effective, so I chose to use this method on the other types of models I trained. I then trained a CART model and a random forest model. The best performing models were the logistic regression model and the random forest model. Using confusion matrices I was able to determine that the LRM correctly guessed 59.176% of the yeses on the training data and 54.704% of the yeses on the test data. The RF model achieved a 59.010% success rate on the training data and 53.704% on the test data.

I was unable to get the confusion matrix to work correctly when I used cross validation with the logistic regression model. I think the data set is large enough this is not that big of a deal. I was able to use cross fold validation with the random forest model. Both models did a pretty good job of predicting positive matches. For something that happens 16% of the time being able to predict that correctly over half the time seems like a pretty good first step.

If I had more time to devote to this project, I would start by taking more time to examine all the variables. There must be some that were left on the cutting room floor that could be useful for training the models. Also I would like to examine more kinds of models. SMOTE, for example, is an oversampling technique that might be a good fit that I haven’t covered in my coursework yet. I would also attempt to create an ensemble of the models so that the strengths of the various models could be magnified, and their weaknesses diminished.

**Appendix 1**

Variables:

iid: unique subject number, group(wave id gender)

id: subject number within wave

gender: Female=0, Male=1

idg: subject number within gender, group(id gender)

condtn: 1=limited choice, 2=extensive choice

round: number of people that met in wave

position: station number where met partner

positin1: station number where started

order: the number of date that night when met partner

partner: partner’s id number the night of event

pid: partner’s iid number

match: 1=yes, 0=no

int\_corr: correlation between participant’s and partner’s ratings of interests in Time

samerace: participant and the partner were the same race. 1= yes, 0=no

age\_o: age of partner

race\_o: race of partner

pf\_o\_att: partner’s stated preference at Time 1 (attr1\_1) for all 6 attributes

dec\_o: decision of partner the night of event

attr\_o: rating by partner the night of the event, for all 6 attributes

round: number of people that met in wave

position: station number where met partner

positin1: station number where started

order: the number of date that night when met partner

partner: partner’s id number the night of event

pid: partner’s iid number

int\_corr: correlation between participant’s and partner’s ratings of interests in Time 1

samerace: participant and the partner were the same race. 1= yes, 0=no

age\_o: age of partner

race\_o: race of partner

pf\_o\_att: partner’s stated preference at Time 1 (attr1\_1) for all 6 attributes

dec\_o: decision of partner the night of event

attr\_o: rating by partner the night of the event, for all 6 attributes

Initial Survey:

age: age of respondent

field: field of study

field\_cd: field coded

1= Law, 2= Math, 3= Social Science, Psychologist, 4= Medical Science, Pharmaceuticals, and Bio Tech, 5= Engineering, 6= English/Creative Writing/ Journalism, 7=History/Religion/Philosophy, 8= Business/Econ/Finance, 9= Education, Academia, 10= Biological Sciences/Chemistry/Physics, 11= Social Work, 12= Undergrad/undecided, 13=Political Science/International Affairs, 14=Film, 15=Fine Arts/Arts Administration, 16=Languages, 17=Architecture, 18=Other

undergrd: school attended for undergraduate degree

mn\_sat: Median SAT score for the undergraduate institution where attended. Taken from Barron’s 25th Edition college profile book. Proxy for intelligence.

tuition: Tuition listed for each response to undergrad in Barron’s 25th Edition college profile book.

race: Black/African American=1, European/Caucasian-American=2, Latino/Hispanic American=3, Asian/Pacific Islander/Asian-American=4, Native American=5,

Other=6

imprace: How important is it to you (on a scale of 1-10) that a person you date be of the same racial/ethnic background?

imprelig: How important is it to you (on a scale of 1-10) that a person you date be of the same religious background?

 from: Where are you from originally (before coming to Columbia)?

zipcode: What was the zip code of the area where you grew up?

income: Median household income based on zipcode using the Census Bureau website:

<http://venus.census.gov/cdrom/lookup/CMD=LIST/DB=C90STF3B/LEV=ZIP>

When there is no income it means that they are either from abroad or did not enter their zip code.

goal: What is your primary goal in participating in this event?

Seemed like a fun night out=1, To meet new people=2, To get a date=3, Looking for a serious relationship=4, To say I did it=5, Other=6

date: In general, how frequently do you go on dates?

Several times a week=1, Twice a week=2, Once a week=3, Twice a month=4, Once a month=5, Several times a year=6, Almost never=7

go out: How often do you go out (not necessarily on dates)?

Ex

career: What is your intended career?

career\_c: career coded

1= Lawyer, 2= Academic/Research, 3= Psychologist, 4= Doctor/Medicine, 5=Engineer, 6= Creative Arts/Entertainment, 7=Banking/Consulting/Finance/Marketing/Business/CEO/-Entrepreneur/Admin, 8= Real Estate, 9= International/Humanitarian Affairs, 10= Undecided, 11=Social Work, 12=Speech Pathology, 13=Politics, 14=Pro sports/Athletics, 15=Other, 16=Journalism, 17=Architecture

12. How interested are you in the following activities, on a scale of 1-10?

sports: Playing sports/ athletics

tvsports: Watching sports

excersice: Body building/exercising

dining: Dining out

museums: Museums/galleries

art: Art

hiking: Hiking/camping

gaming: Gaming

clubbing: Dancing/clubbing

reading: Reading

tv: Watching TV

theater: Theater

movies: Moviesma

concerts: Going to concerts

music: Music

shopping: Shopping

yoga: Yoga/meditation

exphappy: Overall, on a scale of 1-10, how happy do you expect to be with the people you meet during the speed-dating event?

expnum: Out of the 20 people you will meet, how many do you expect will be interested in dating you?

We want to know what you look for in the opposite sex.

Waves 6-9: Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 1-5, 10-21: You have 100 points to distribute among the following attributes -- give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100.

attr1\_1:  Attractive

sinc1\_1: Sincere

intel1\_1: Intelligent

fun1\_1: Fun

amb1\_1: Ambitious

shar1\_1: Has shared interests/hobbies

Now we want to know what you think MOST of your fellow men/women look for in the opposite sex.

Waves 6-9: Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 10-21 : You have 100 points to distribute among the following attributes -- give more points to those attributes that you think your fellow men/women find more important in a potential date and fewer points to those attributes that they find less important in a potential date. Total points must equal 100.

attr4\_1: Attractive

sinc4\_1: Sincere

intel4\_1: Intelligent

fun4\_1: Fun

amb4\_1: Ambitious

shar4\_1: Shared Interests/Hobbies

What do you think the opposite sex looks for in a date?

Waves 6-9: Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

 Waves 1-5 and 10-21: Please distribute 100 points among the following attributes -- give more points to those attributes that you think are more important to members of the opposite sex when they are deciding whether to date someone. Total points must equal 100.

attr2\_1: Attractive

sinc2\_1: Sincere

int2\_1: Intelligent

fun2\_1: Fun

amb2\_1: Ambitious

shar2\_1: Has shared interests/hobbies

How do you think you measure up?

Please rate your opinion of your own attributes, on a scale of 1-10 (be honest!):

attr3\_1 : Attractive

sinc3\_1: Sincere

int3\_1: Intelligent

fun3\_1: Fun

amb3\_1: Ambitious

And finally, how do you think others perceive you?

Please rate yourself how you think others would rate you on each of the following attributes, on a scale of 1-10 (1=awful, 10=great)

attr5\_1 : Attractive

sinc5\_1: Sincere

int5\_1: Intelligent

fun5\_1: Fun

amb5\_1: Ambitious

Half way through meeting all potential dates during the night of the event on their scorecard:

Please rate the importance of the following attributes in a potential date on a scale of 1-10: (1=not at all important, 10=extremely important).

attr1\_s: Attractive

sinc1\_s: Sincere

intel1\_s: Intelligent

fun1\_s: Fun

amb1\_s: Ambitious

shar1\_s: Shared Interests/Hobbies

Please rate your opinion of your own attributes, on a scale of 1-10 (1=awful, 10=great) --Be honest!

attr3\_s: Attractive

sinc3\_s: Sincere

intel3\_s: Intelligent

fun3\_s: Fun

amb3\_s: Ambitious

Followup/Time2:

[Survey is filled out the day after participating in the event. Subjects must have submitted this in order to be sent their matches.]

satis\_2: Overall, how satisfied were you with the people you met? (1=not at all satisfied, 10=extremely satisfied)

length: Four minutes is: Too little=1, Too much=2, Just Right=3

numdat\_2: The number of Speed "Dates" you had was:

Too few=1, Too many=2, Just right=3

Now, think back to your yes/no decisions during the Speed Dating event. Try to distribute the 100 points among these six attributes in the way that best reflects the actual importance of these attributes in your decisions. Give more points to those attributes that were more important in your decisions, and fewer points to those attributes that were less important in your decisions. Total points must equal 100.

attr7\_2: Attractive

sinc7\_2: Sincere

intel7\_2: Intelligent

fun7\_2: Fun

amb7\_2: Ambitious

shar7\_2: Has shared interests/hobbies

We want to know what you look for in the opposite sex.

Waves 1-5 and 10-21: You have 100 points to distribute among the following attributes -- give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100.

Waves 6-9: Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important):

attr1\_2 : Attractive

sinc1\_2: Sincere

intel1\_2: Intelligent

fun1\_2: Fun

amb1\_2: Ambitious

shar1\_2: Has shared interests/hobbies

What do you think MOST of your fellow men/women look for in the opposite sex?

You have 100 points to distribute among the following attributes -- give more points to those attributes that you think your fellow men/women find more important in a potential date, and fewer points to those attributes that they find less important in a potential date.

Total points must equal 100.

attr4\_2: Attractive

sinc4\_2: Sincere

intel4\_2: Intelligent

fun4\_2: Fun

amb4\_2: Ambitious

shar4\_2: Shared Interests/Hobbies

What do you think the opposite sex looks for in a date?

Please distribute 100 points among the following attributes -- give more points to those attributes that you think are more important to members of the opposite sex when they are deciding whether to date someone. Total points must equal 100.

attr2\_2 : Attractive

sinc2\_2: Sincere

intel2\_2: Intelligent

fun2\_2: Fun

amb2\_2: Ambitious

shar2\_2: Has shared interests/hobbies

How do you think you measure up?

Please rate your opinion of your own attributes, on a scale of 1-10 (1= awful and 10=great). Be honest!

attr3\_2 : Attractive

sinc3\_2: Sincere

int3\_2: Intelligent

fun3\_2: Fun

amb3\_2: Ambitious

And finally, how do you think others perceive you?

Please rate yourself how you think others would rate you on each of the following attributes, on a scale of 1-10 (1=awful, 10=great)

attr5\_2 : Attractive

sinc5\_2: Sincere

int5\_2: Intelligent

fun5\_2: Fun

amb5\_2: Ambitious

followup2/ Time3:

[Subjects filled out 3-4 weeks after they had been sent their matches]

SINCE HURRYDATING…

1. Of the matches that you received:

you\_call: (a) How many have you contacted to set up a date?

them\_cal: (b) How many have contacted you?

date\_3: Have you been on a date with any of your matches?

Yes=1, No=2

If you have been on at least one date, please answer the following:

numdat\_3: (a) How many of your matches have you been on a date with so far?

num\_in\_3: If yes, how many?

What do you look for in the opposite sex?

Please distribute 100 points among the following attributes -- give more to attributes that were more important in your decisions when Hurrydating, and less to attributes that were less important. Total points must equal 100.

We want to know what you look for in the opposite sex.

Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important):

attr1\_3 : Attractive

sinc1\_3: Sincere

intel1\_3: Intelligent

fun1\_3: Fun

amb1\_3: Ambitious

shar1\_3: Has shared interests/hobbies

Now, think back to your yes/no decisions during the night of the Speed Dating event. Try to distribute the 100 points among these six attributes in the way that best reflects the actual importance of these attributes in your decisions. Give more points to those attributes that were more important in your decisions, and fewer points to those attributes that less important in your decisions. Total points must equal 100.

attr7\_3 : Attractive

sinc7\_3: Sincere

intel7\_3: Intelligent

fun7\_3: Fun

amb7\_3: Ambitious

shar7\_3: Has shared interests/hobbies

Now we want to know what you think MOST of your fellow men/women look for in the opposite sex.

Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

attr4\_3 : Attractive

sinc4\_3: Sincere

intel4\_3: Intelligent

fun4\_3: Fun

amb4\_3: Ambitious

shar4\_3: Has shared interests/hobbies

What do you think the opposite sex looks for in a date?

Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

attr2\_3 : Attractive

sinc2\_3: Sincere

intel2\_3: Intelligent

fun2\_3: Fun

amb2\_3: Ambitious

share2\_3: Has shared interests/hobbies

Please rate your opinion of your own attributes, on a scale of 1-10 (1= awful and 10=great).

attr3 \_3: Attractive

sinc3\_3: Sincere

intel3\_3: Intelligent

fun3\_3: Fun

amb3\_3: Ambitious

And finally, how do you think others perceive you?

Please rate yourself how you think others would rate you on each of the following attributes, on a scale of 1-10 (1=awful, 10=great)

attr5\_3 : Attractive

sinc5\_3: Sincere

int5\_3: Intelligent

fun5\_3: Fun

amb5\_3: Ambitious

**Appendix 2**

DatingData.R

serio

Sun Apr 28 03:48:50 2019

#Check out my packages

require(ISLR)

## Loading required package: ISLR

## Warning: package 'ISLR' was built under R version 3.5.2

require(boot)

## Loading required package: boot

require(caret)

## Loading required package: caret

## Warning: package 'caret' was built under R version 3.5.3

## Loading required package: lattice

##

## Attaching package: 'lattice'

## The following object is masked from 'package:boot':

##

## melanoma

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.5.2

require(ggplot2)

require(plyr)

## Loading required package: plyr

## Warning: package 'plyr' was built under R version 3.5.1

require(ROCR)

## Loading required package: ROCR

## Warning: package 'ROCR' was built under R version 3.5.2

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.5.2

##

## Attaching package: 'gplots'

## The following object is masked from 'package:stats':

##

## lowess

require(ranger)

## Loading required package: ranger

## Warning: package 'ranger' was built under R version 3.5.3

#Aquire all the data. This dataset contains 8378 observations of 195 variables.

dating <-read.csv("Speed Dating Data.csv")

# Function that returns Root Mean Squared Error

rmse <- function(error)

{

sqrt(mean(error^2))

}

# Function that returns Mean Absolute Error

mae <- function(error)

{

mean(abs(error))

}

#Examine the dataset

summary(dating)

## iid id gender idg

## Min. : 1.0 Min. : 1.00 Min. :0.0000 Min. : 1.00

## 1st Qu.:154.0 1st Qu.: 4.00 1st Qu.:0.0000 1st Qu.: 8.00

## Median :281.0 Median : 8.00 Median :1.0000 Median :16.00

## Mean :283.7 Mean : 8.96 Mean :0.5006 Mean :17.33

## 3rd Qu.:407.0 3rd Qu.:13.00 3rd Qu.:1.0000 3rd Qu.:26.00

## Max. :552.0 Max. :22.00 Max. :1.0000 Max. :44.00

## NA's :1

## condtn wave round position

## Min. :1.000 Min. : 1.00 Min. : 5.00 Min. : 1.000

## 1st Qu.:2.000 1st Qu.: 7.00 1st Qu.:14.00 1st Qu.: 4.000

## Median :2.000 Median :11.00 Median :18.00 Median : 8.000

## Mean :1.829 Mean :11.35 Mean :16.87 Mean : 9.043

## 3rd Qu.:2.000 3rd Qu.:15.00 3rd Qu.:20.00 3rd Qu.:13.000

## Max. :2.000 Max. :21.00 Max. :22.00 Max. :22.000

##

## positin1 order partner pid

## Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 1.0

## 1st Qu.: 4.000 1st Qu.: 4.000 1st Qu.: 4.000 1st Qu.:154.0

## Median : 9.000 Median : 8.000 Median : 8.000 Median :281.0

## Mean : 9.296 Mean : 8.928 Mean : 8.964 Mean :283.9

## 3rd Qu.:14.000 3rd Qu.:13.000 3rd Qu.:13.000 3rd Qu.:408.0

## Max. :22.000 Max. :22.000 Max. :22.000 Max. :552.0

## NA's :1846 NA's :10

## match int\_corr samerace age\_o

## Min. :0.0000 Min. :-0.830 Min. :0.0000 Min. :18.00

## 1st Qu.:0.0000 1st Qu.:-0.020 1st Qu.:0.0000 1st Qu.:24.00

## Median :0.0000 Median : 0.210 Median :0.0000 Median :26.00

## Mean :0.1647 Mean : 0.196 Mean :0.3958 Mean :26.36

## 3rd Qu.:0.0000 3rd Qu.: 0.430 3rd Qu.:1.0000 3rd Qu.:28.00

## Max. :1.0000 Max. : 0.910 Max. :1.0000 Max. :55.00

## NA's :158 NA's :104

## race\_o pf\_o\_att pf\_o\_sin pf\_o\_int

## Min. :1.000 Min. : 0.0 Min. : 0.00 Min. : 0.00

## 1st Qu.:2.000 1st Qu.: 15.0 1st Qu.:15.00 1st Qu.:17.39

## Median :2.000 Median : 20.0 Median :18.37 Median :20.00

## Mean :2.757 Mean : 22.5 Mean :17.40 Mean :20.27

## 3rd Qu.:4.000 3rd Qu.: 25.0 3rd Qu.:20.00 3rd Qu.:23.81

## Max. :6.000 Max. :100.0 Max. :60.00 Max. :50.00

## NA's :73 NA's :89 NA's :89 NA's :89

## pf\_o\_fun pf\_o\_amb pf\_o\_sha dec\_o

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. :0.0000

## 1st Qu.:15.00 1st Qu.: 5.00 1st Qu.: 9.52 1st Qu.:0.0000

## Median :18.00 Median :10.00 Median :10.64 Median :0.0000

## Mean :17.46 Mean :10.69 Mean :11.85 Mean :0.4196

## 3rd Qu.:20.00 3rd Qu.:15.00 3rd Qu.:16.00 3rd Qu.:1.0000

## Max. :50.00 Max. :53.00 Max. :30.00 Max. :1.0000

## NA's :98 NA's :107 NA's :129

## attr\_o sinc\_o intel\_o fun\_o

## Min. : 0.00 Min. : 0.000 Min. : 0.000 Min. : 0.000

## 1st Qu.: 5.00 1st Qu.: 6.000 1st Qu.: 6.000 1st Qu.: 5.000

## Median : 6.00 Median : 7.000 Median : 7.000 Median : 7.000

## Mean : 6.19 Mean : 7.175 Mean : 7.369 Mean : 6.401

## 3rd Qu.: 8.00 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 8.000

## Max. :10.50 Max. :10.000 Max. :10.000 Max. :11.000

## NA's :212 NA's :287 NA's :306 NA's :360

## amb\_o shar\_o like\_o prob\_o

## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000

## 1st Qu.: 6.000 1st Qu.: 4.000 1st Qu.: 5.000 1st Qu.: 4.000

## Median : 7.000 Median : 6.000 Median : 6.000 Median : 5.000

## Mean : 6.778 Mean : 5.475 Mean : 6.135 Mean : 5.208

## 3rd Qu.: 8.000 3rd Qu.: 7.000 3rd Qu.: 7.000 3rd Qu.: 7.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :722 NA's :1076 NA's :250 NA's :318

## met\_o age field

## Min. :1.00 Min. :18.00 Business : 521

## 1st Qu.:2.00 1st Qu.:24.00 MBA : 468

## Median :2.00 Median :26.00 Law : 462

## Mean :1.96 Mean :26.36 Social Work : 378

## 3rd Qu.:2.00 3rd Qu.:28.00 International Affairs : 252

## Max. :8.00 Max. :55.00 Electrical Engineering: 164

## NA's :385 NA's :95 (Other) :6133

## field\_cd undergra mn\_sat tuition

## Min. : 1.000 :3464 :5245 :4795

## 1st Qu.: 5.000 UC Berkeley: 107 1,400.00: 403 26,908.00: 241

## Median : 8.000 Harvard : 104 1,430.00: 262 26,019.00: 174

## Mean : 7.662 Columbia : 95 1,290.00: 190 15,162.00: 138

## 3rd Qu.:10.000 Yale : 86 1,450.00: 163 25,380.00: 112

## Max. :18.000 NYU : 78 1,340.00: 146 26,062.00: 108

## NA's :82 (Other) :4444 (Other) :1969 (Other) :2810

## race imprace imprelig from

## Min. :1.000 Min. : 0.000 Min. : 1.000 New York : 522

## 1st Qu.:2.000 1st Qu.: 1.000 1st Qu.: 1.000 New Jersey : 365

## Median :2.000 Median : 3.000 Median : 3.000 California : 301

## Mean :2.757 Mean : 3.785 Mean : 3.652 China : 139

## 3rd Qu.:4.000 3rd Qu.: 6.000 3rd Qu.: 6.000 Italy : 132

## Max. :6.000 Max. :10.000 Max. :10.000 New York City: 130

## NA's :63 NA's :79 NA's :79 (Other) :6789

## zipcode income goal date

## :1064 :4099 Min. :1.000 Min. :1.000

## 0 : 355 55,080.00: 124 1st Qu.:1.000 1st Qu.:4.000

## 10,021 : 139 53,229.00: 41 Median :2.000 Median :5.000

## 10,027 : 128 25,401.00: 39 Mean :2.122 Mean :5.007

## 10,025 : 121 33,772.00: 37 3rd Qu.:2.000 3rd Qu.:6.000

## 19,087 : 48 49,409.00: 37 Max. :6.000 Max. :7.000

## (Other):6523 (Other) :4001 NA's :79 NA's :97

## go\_out career career\_c sports

## Min. :1.000 Finance : 202 Min. : 1.000 Min. : 1.000

## 1st Qu.:1.000 professor : 199 1st Qu.: 2.000 1st Qu.: 4.000

## Median :2.000 Lawyer : 154 Median : 6.000 Median : 7.000

## Mean :2.158 Professor : 148 Mean : 5.278 Mean : 6.425

## 3rd Qu.:3.000 Consulting : 147 3rd Qu.: 7.000 3rd Qu.: 9.000

## Max. :7.000 Social Worker: 136 Max. :17.000 Max. :10.000

## NA's :79 (Other) :7392 NA's :138 NA's :79

## tvsports exercise dining museums

## Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 0.000

## 1st Qu.: 2.000 1st Qu.: 5.000 1st Qu.: 7.000 1st Qu.: 6.000

## Median : 4.000 Median : 6.000 Median : 8.000 Median : 7.000

## Mean : 4.575 Mean : 6.246 Mean : 7.784 Mean : 6.986

## 3rd Qu.: 7.000 3rd Qu.: 8.000 3rd Qu.: 9.000 3rd Qu.: 9.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :79 NA's :79 NA's :79 NA's :79

## art hiking gaming clubbing

## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000

## 1st Qu.: 5.000 1st Qu.: 4.000 1st Qu.: 2.000 1st Qu.: 4.000

## Median : 7.000 Median : 6.000 Median : 3.000 Median : 6.000

## Mean : 6.715 Mean : 5.737 Mean : 3.881 Mean : 5.746

## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 6.000 3rd Qu.: 8.000

## Max. :10.000 Max. :10.000 Max. :14.000 Max. :10.000

## NA's :79 NA's :79 NA's :79 NA's :79

## reading tv theater movies

## Min. : 1.000 Min. : 1.000 Min. : 0.000 Min. : 0.00

## 1st Qu.: 7.000 1st Qu.: 3.000 1st Qu.: 5.000 1st Qu.: 7.00

## Median : 8.000 Median : 6.000 Median : 7.000 Median : 8.00

## Mean : 7.679 Mean : 5.304 Mean : 6.776 Mean : 7.92

## 3rd Qu.: 9.000 3rd Qu.: 7.000 3rd Qu.: 9.000 3rd Qu.: 9.00

## Max. :13.000 Max. :10.000 Max. :10.000 Max. :10.00

## NA's :79 NA's :79 NA's :79 NA's :79

## concerts music shopping yoga

## Min. : 0.000 Min. : 1.000 Min. : 1.000 Min. : 0.000

## 1st Qu.: 5.000 1st Qu.: 7.000 1st Qu.: 4.000 1st Qu.: 2.000

## Median : 7.000 Median : 8.000 Median : 6.000 Median : 4.000

## Mean : 6.825 Mean : 7.851 Mean : 5.631 Mean : 4.339

## 3rd Qu.: 8.000 3rd Qu.: 9.000 3rd Qu.: 8.000 3rd Qu.: 7.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :79 NA's :79 NA's :79 NA's :79

## exphappy expnum attr1\_1 sinc1\_1

## Min. : 1.000 Min. : 0.000 Min. : 0.00 Min. : 0.00

## 1st Qu.: 5.000 1st Qu.: 2.000 1st Qu.: 15.00 1st Qu.:15.00

## Median : 6.000 Median : 4.000 Median : 20.00 Median :18.18

## Mean : 5.534 Mean : 5.571 Mean : 22.51 Mean :17.40

## 3rd Qu.: 7.000 3rd Qu.: 8.000 3rd Qu.: 25.00 3rd Qu.:20.00

## Max. :10.000 Max. :20.000 Max. :100.00 Max. :60.00

## NA's :101 NA's :6578 NA's :79 NA's :79

## intel1\_1 fun1\_1 amb1\_1 shar1\_1

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.:17.39 1st Qu.:15.00 1st Qu.: 5.00 1st Qu.: 9.52

## Median :20.00 Median :18.00 Median :10.00 Median :10.64

## Mean :20.27 Mean :17.46 Mean :10.68 Mean :11.85

## 3rd Qu.:23.81 3rd Qu.:20.00 3rd Qu.:15.00 3rd Qu.:16.00

## Max. :50.00 Max. :50.00 Max. :53.00 Max. :30.00

## NA's :79 NA's :89 NA's :99 NA's :121

## attr4\_1 sinc4\_1 intel4\_1 fun4\_1

## Min. : 5.00 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.:10.00 1st Qu.: 6.00 1st Qu.: 8.00 1st Qu.:10.00

## Median :25.00 Median :10.00 Median :10.00 Median :15.00

## Mean :26.39 Mean :11.07 Mean :12.64 Mean :15.57

## 3rd Qu.:35.00 3rd Qu.:15.00 3rd Qu.:16.00 3rd Qu.:20.00

## Max. :95.00 Max. :35.00 Max. :35.00 Max. :45.00

## NA's :1889 NA's :1889 NA's :1889 NA's :1889

## amb4\_1 shar4\_1 attr2\_1 sinc2\_1

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.: 5.00 1st Qu.: 7.00 1st Qu.: 20.00 1st Qu.:10.00

## Median :10.00 Median :10.00 Median : 25.00 Median :15.00

## Mean : 9.78 Mean :11.01 Mean : 30.36 Mean :13.27

## 3rd Qu.:15.00 3rd Qu.:15.00 3rd Qu.: 40.00 3rd Qu.:18.75

## Max. :50.00 Max. :40.00 Max. :100.00 Max. :50.00

## NA's :1889 NA's :1911 NA's :79 NA's :79

## intel2\_1 fun2\_1 amb2\_1 shar2\_1

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.:10.00 1st Qu.:15.00 1st Qu.: 6.00 1st Qu.:10.00

## Median :15.00 Median :20.00 Median :10.00 Median :10.00

## Mean :14.42 Mean :18.42 Mean :11.74 Mean :11.85

## 3rd Qu.:20.00 3rd Qu.:20.00 3rd Qu.:15.00 3rd Qu.:15.63

## Max. :40.00 Max. :50.00 Max. :50.00 Max. :30.00

## NA's :79 NA's :79 NA's :89 NA's :89

## attr3\_1 sinc3\_1 fun3\_1 intel3\_1

## Min. : 2.000 Min. : 2.000 Min. : 2.000 Min. : 3.000

## 1st Qu.: 6.000 1st Qu.: 8.000 1st Qu.: 7.000 1st Qu.: 8.000

## Median : 7.000 Median : 8.000 Median : 8.000 Median : 8.000

## Mean : 7.085 Mean : 8.295 Mean : 7.704 Mean : 8.404

## 3rd Qu.: 8.000 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 9.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :105 NA's :105 NA's :105 NA's :105

## amb3\_1 attr5\_1 sinc5\_1 intel5\_1

## Min. : 2.000 Min. : 2.000 Min. : 1.000 Min. : 3.000

## 1st Qu.: 7.000 1st Qu.: 6.000 1st Qu.: 7.000 1st Qu.: 8.000

## Median : 8.000 Median : 7.000 Median : 8.000 Median : 8.000

## Mean : 7.578 Mean : 6.942 Mean : 7.927 Mean : 8.284

## 3rd Qu.: 9.000 3rd Qu.: 8.000 3rd Qu.: 9.000 3rd Qu.: 9.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :105 NA's :3472 NA's :3472 NA's :3472

## fun5\_1 amb5\_1 dec attr

## Min. : 2.000 Min. : 1.000 Min. :0.0000 Min. : 0.00

## 1st Qu.: 6.000 1st Qu.: 7.000 1st Qu.:0.0000 1st Qu.: 5.00

## Median : 8.000 Median : 8.000 Median :0.0000 Median : 6.00

## Mean : 7.426 Mean : 7.618 Mean :0.4199 Mean : 6.19

## 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.:1.0000 3rd Qu.: 8.00

## Max. :10.000 Max. :10.000 Max. :1.0000 Max. :10.00

## NA's :3472 NA's :3472 NA's :202

## sinc intel fun amb

## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000

## 1st Qu.: 6.000 1st Qu.: 6.000 1st Qu.: 5.000 1st Qu.: 6.000

## Median : 7.000 Median : 7.000 Median : 7.000 Median : 7.000

## Mean : 7.175 Mean : 7.369 Mean : 6.401 Mean : 6.777

## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 8.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :277 NA's :296 NA's :350 NA's :712

## shar like prob met

## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. :0.0000

## 1st Qu.: 4.000 1st Qu.: 5.000 1st Qu.: 4.000 1st Qu.:0.0000

## Median : 6.000 Median : 6.000 Median : 5.000 Median :0.0000

## Mean : 5.475 Mean : 6.134 Mean : 5.208 Mean :0.9488

## 3rd Qu.: 7.000 3rd Qu.: 7.000 3rd Qu.: 7.000 3rd Qu.:2.0000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :8.0000

## NA's :1067 NA's :240 NA's :309 NA's :375

## match\_es attr1\_s sinc1\_s intel1\_s

## Min. : 0.000 Min. : 3.00 Min. : 0.00 Min. : 0.00

## 1st Qu.: 2.000 1st Qu.:14.81 1st Qu.:10.00 1st Qu.:10.00

## Median : 3.000 Median :17.65 Median :15.79 Median :18.42

## Mean : 3.208 Mean :20.79 Mean :15.43 Mean :17.24

## 3rd Qu.: 4.000 3rd Qu.:25.00 3rd Qu.:20.00 3rd Qu.:20.00

## Max. :18.000 Max. :95.00 Max. :50.00 Max. :40.00

## NA's :1173 NA's :4282 NA's :4282 NA's :4282

## fun1\_s amb1\_s shar1\_s attr3\_s

## Min. : 1.00 Min. : 0.00 Min. : 0.00 Min. : 3.000

## 1st Qu.:10.00 1st Qu.: 7.00 1st Qu.: 9.00 1st Qu.: 7.000

## Median :15.91 Median :10.00 Median :12.50 Median : 7.000

## Mean :15.26 Mean :11.14 Mean :12.46 Mean : 7.211

## 3rd Qu.:20.00 3rd Qu.:15.00 3rd Qu.:16.28 3rd Qu.: 8.000

## Max. :40.00 Max. :23.81 Max. :30.00 Max. :10.000

## NA's :4282 NA's :4282 NA's :4282 NA's :4378

## sinc3\_s intel3\_s fun3\_s amb3\_s

## Min. : 1.000 Min. : 4.000 Min. : 3.000 Min. : 2.000

## 1st Qu.: 7.000 1st Qu.: 8.000 1st Qu.: 7.000 1st Qu.: 7.000

## Median : 8.000 Median : 8.000 Median : 8.000 Median : 8.000

## Mean : 8.082 Mean : 8.258 Mean : 7.692 Mean : 7.589

## 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 9.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :4378 NA's :4378 NA's :4378 NA's :4378

## satis\_2 length numdat\_2 attr7\_2

## Min. : 1.000 Min. :1.000 Min. :1.000 Min. :10.00

## 1st Qu.: 5.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:20.00

## Median : 6.000 Median :1.000 Median :2.000 Median :30.00

## Mean : 5.712 Mean :1.843 Mean :2.338 Mean :32.82

## 3rd Qu.: 7.000 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:40.00

## Max. :10.000 Max. :3.000 Max. :3.000 Max. :80.00

## NA's :915 NA's :915 NA's :945 NA's :6394

## sinc7\_2 intel7\_2 fun7\_2 amb7\_2

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.000

## 1st Qu.:10.00 1st Qu.:10.00 1st Qu.:10.00 1st Qu.: 0.000

## Median :10.00 Median :15.00 Median :20.00 Median : 5.000

## Mean :13.53 Mean :15.29 Mean :18.87 Mean : 7.287

## 3rd Qu.:20.00 3rd Qu.:20.00 3rd Qu.:24.00 3rd Qu.:10.000

## Max. :40.00 Max. :50.00 Max. :50.00 Max. :20.000

## NA's :6423 NA's :6394 NA's :6394 NA's :6423

## shar7\_2 attr1\_2 sinc1\_2 intel1\_2

## Min. : 0.00 Min. : 5.00 Min. : 0.00 Min. : 0.00

## 1st Qu.: 5.00 1st Qu.:16.67 1st Qu.:10.00 1st Qu.:15.00

## Median :10.00 Median :20.00 Median :16.67 Median :19.05

## Mean :12.16 Mean :26.22 Mean :15.87 Mean :17.81

## 3rd Qu.:20.00 3rd Qu.:30.00 3rd Qu.:20.00 3rd Qu.:20.00

## Max. :40.00 Max. :85.00 Max. :50.00 Max. :40.00

## NA's :6404 NA's :933 NA's :915 NA's :915

## fun1\_2 amb1\_2 shar1\_2 attr4\_2

## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. : 6.00

## 1st Qu.:15.00 1st Qu.: 5.000 1st Qu.:10.00 1st Qu.: 10.00

## Median :18.37 Median :10.000 Median :13.00 Median : 25.00

## Mean :17.65 Mean : 9.913 Mean :12.76 Mean : 26.81

## 3rd Qu.:20.00 3rd Qu.:15.000 3rd Qu.:16.67 3rd Qu.: 40.00

## Max. :50.00 Max. :22.220 Max. :35.00 Max. :100.00

## NA's :915 NA's :915 NA's :915 NA's :2603

## sinc4\_2 intel4\_2 fun4\_2 amb4\_2

## Min. : 0.00 Min. : 0.0 Min. : 0.00 Min. : 0.000

## 1st Qu.: 8.00 1st Qu.: 8.0 1st Qu.: 9.00 1st Qu.: 5.000

## Median :10.00 Median :10.0 Median :15.00 Median :10.000

## Mean :11.93 Mean :12.1 Mean :15.16 Mean : 9.342

## 3rd Qu.:15.00 3rd Qu.:15.0 3rd Qu.:20.00 3rd Qu.:10.000

## Max. :35.00 Max. :40.0 Max. :50.00 Max. :35.000

## NA's :2603 NA's :2603 NA's :2603 NA's :2603

## shar4\_2 attr2\_2 sinc2\_2 intel2\_2

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.: 7.00 1st Qu.:19.15 1st Qu.:10.00 1st Qu.:10.00

## Median :10.00 Median :25.00 Median :15.00 Median :15.00

## Mean :11.32 Mean :29.34 Mean :13.90 Mean :13.96

## 3rd Qu.:15.00 3rd Qu.:38.46 3rd Qu.:19.23 3rd Qu.:17.39

## Max. :40.00 Max. :85.00 Max. :40.00 Max. :30.77

## NA's :2603 NA's :2603 NA's :2603 NA's :2603

## fun2\_2 amb2\_2 shar2\_2 attr3\_2

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 2.000

## 1st Qu.:15.00 1st Qu.:10.00 1st Qu.:10.00 1st Qu.: 7.000

## Median :18.52 Median :10.00 Median :13.95 Median : 7.000

## Mean :17.97 Mean :11.91 Mean :12.89 Mean : 7.125

## 3rd Qu.:20.00 3rd Qu.:15.09 3rd Qu.:16.52 3rd Qu.: 8.000

## Max. :40.00 Max. :50.00 Max. :30.00 Max. :10.000

## NA's :2603 NA's :2603 NA's :2603 NA's :915

## sinc3\_2 intel3\_2 fun3\_2 amb3\_2

## Min. : 2.000 Min. : 4.000 Min. : 1.000 Min. : 2.000

## 1st Qu.: 7.000 1st Qu.: 8.000 1st Qu.: 7.000 1st Qu.: 7.000

## Median : 8.000 Median : 8.000 Median : 8.000 Median : 8.000

## Mean : 7.931 Mean : 8.239 Mean : 7.602 Mean : 7.487

## 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 9.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :915 NA's :915 NA's :915 NA's :915

## attr5\_2 sinc5\_2 intel5\_2 fun5\_2

## Min. : 2.000 Min. : 2.000 Min. : 2.000 Min. : 2.000

## 1st Qu.: 6.000 1st Qu.: 6.000 1st Qu.: 7.000 1st Qu.: 6.000

## Median : 7.000 Median : 8.000 Median : 8.000 Median : 7.000

## Mean : 6.828 Mean : 7.394 Mean : 7.839 Mean : 7.279

## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 9.000 3rd Qu.: 8.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :4001 NA's :4001 NA's :4001 NA's :4001

## amb5\_2 you\_call them\_cal date\_3

## Min. : 2.000 Min. : 0.000 Min. :0.000 Min. :0.000

## 1st Qu.: 6.000 1st Qu.: 0.000 1st Qu.:0.000 1st Qu.:0.000

## Median : 7.000 Median : 0.000 Median :1.000 Median :0.000

## Mean : 7.332 Mean : 0.781 Mean :0.982 Mean :0.377

## 3rd Qu.: 8.000 3rd Qu.: 1.000 3rd Qu.:1.000 3rd Qu.:1.000

## Max. :10.000 Max. :21.000 Max. :9.000 Max. :1.000

## NA's :4001 NA's :4404 NA's :4404 NA's :4404

## numdat\_3 num\_in\_3 attr1\_3 sinc1\_3

## Min. :0.000 Min. :0.000 Min. : 0.00 Min. : 0.00

## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:15.22 1st Qu.:10.00

## Median :1.000 Median :1.000 Median :20.00 Median :16.67

## Mean :1.231 Mean :0.934 Mean :24.39 Mean :16.59

## 3rd Qu.:1.000 3rd Qu.:1.000 3rd Qu.:30.00 3rd Qu.:20.00

## Max. :9.000 Max. :4.000 Max. :80.00 Max. :65.00

## NA's :6882 NA's :7710 NA's :4404 NA's :4404

## intel1\_3 fun1\_3 amb1\_3 shar1\_3

## Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 0.00

## 1st Qu.:16.67 1st Qu.:14.81 1st Qu.: 5.0 1st Qu.:10.00

## Median :20.00 Median :16.33 Median :10.0 Median :14.29

## Mean :19.41 Mean :16.23 Mean :10.9 Mean :12.70

## 3rd Qu.:20.00 3rd Qu.:20.00 3rd Qu.:15.0 3rd Qu.:16.67

## Max. :45.00 Max. :30.00 Max. :30.0 Max. :55.00

## NA's :4404 NA's :4404 NA's :4404 NA's :4404

## attr7\_3 sinc7\_3 intel7\_3 fun7\_3

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.:20.00 1st Qu.:10.00 1st Qu.:10.00 1st Qu.:10.00

## Median :25.00 Median :15.00 Median :18.00 Median :17.00

## Mean :31.33 Mean :15.65 Mean :16.68 Mean :16.42

## 3rd Qu.:40.00 3rd Qu.:20.00 3rd Qu.:20.00 3rd Qu.:20.00

## Max. :80.00 Max. :60.00 Max. :45.00 Max. :40.00

## NA's :6362 NA's :6362 NA's :6362 NA's :6362

## amb7\_3 shar7\_3 attr4\_3 sinc4\_3

## Min. : 0.000 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.: 0.000 1st Qu.: 5.00 1st Qu.:10.00 1st Qu.: 7.00

## Median :10.000 Median :10.00 Median :20.00 Median :10.00

## Mean : 7.824 Mean :12.21 Mean :25.61 Mean :10.75

## 3rd Qu.:10.000 3rd Qu.:20.00 3rd Qu.:37.00 3rd Qu.:15.00

## Max. :30.000 Max. :55.00 Max. :80.00 Max. :40.00

## NA's :6362 NA's :6362 NA's :5419 NA's :5419

## intel4\_3 fun4\_3 amb4\_3 shar4\_3

## Min. : 0.00 Min. : 0.00 Min. : 0.000 Min. : 0.00

## 1st Qu.: 7.00 1st Qu.: 9.00 1st Qu.: 5.000 1st Qu.: 7.00

## Median :10.00 Median :12.00 Median : 9.000 Median :10.00

## Mean :11.53 Mean :14.28 Mean : 9.208 Mean :11.25

## 3rd Qu.:15.00 3rd Qu.:20.00 3rd Qu.:10.000 3rd Qu.:15.00

## Max. :30.00 Max. :30.00 Max. :40.000 Max. :45.00

## NA's :5419 NA's :5419 NA's :5419 NA's :5419

## attr2\_3 sinc2\_3 intel2\_3 fun2\_3

## Min. : 5.00 Min. : 0.00 Min. : 0.00 Min. : 0.00

## 1st Qu.:10.00 1st Qu.: 7.00 1st Qu.: 7.00 1st Qu.: 9.00

## Median :20.00 Median :10.00 Median :10.00 Median :15.00

## Mean :24.97 Mean :10.92 Mean :11.95 Mean :14.96

## 3rd Qu.:35.00 3rd Qu.:15.00 3rd Qu.:15.00 3rd Qu.:20.00

## Max. :80.00 Max. :50.00 Max. :60.00 Max. :40.00

## NA's :5419 NA's :5419 NA's :5419 NA's :5419

## amb2\_3 shar2\_3 attr3\_3 sinc3\_3

## Min. : 0.000 Min. : 0.00 Min. : 2.00 Min. : 2.000

## 1st Qu.: 6.000 1st Qu.: 5.00 1st Qu.: 7.00 1st Qu.: 7.000

## Median :10.000 Median :10.00 Median : 7.00 Median : 8.000

## Mean : 9.526 Mean :11.97 Mean : 7.24 Mean : 8.093

## 3rd Qu.:10.000 3rd Qu.:15.00 3rd Qu.: 8.00 3rd Qu.: 9.000

## Max. :50.000 Max. :45.00 Max. :12.00 Max. :12.000

## NA's :5419 NA's :6362 NA's :4404 NA's :4404

## intel3\_3 fun3\_3 amb3\_3 attr5\_3

## Min. : 3.000 Min. : 2.000 Min. : 1.000 Min. : 2.00

## 1st Qu.: 8.000 1st Qu.: 7.000 1st Qu.: 6.000 1st Qu.: 6.00

## Median : 8.000 Median : 8.000 Median : 8.000 Median : 7.00

## Mean : 8.389 Mean : 7.659 Mean : 7.392 Mean : 6.81

## 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 8.00

## Max. :12.000 Max. :12.000 Max. :12.000 Max. :10.00

## NA's :4404 NA's :4404 NA's :4404 NA's :6362

## sinc5\_3 intel5\_3 fun5\_3 amb5\_3

## Min. : 2.000 Min. : 4.000 Min. : 1.000 Min. : 1.000

## 1st Qu.: 7.000 1st Qu.: 7.000 1st Qu.: 6.000 1st Qu.: 6.000

## Median : 8.000 Median : 8.000 Median : 7.000 Median : 7.000

## Mean : 7.615 Mean : 7.933 Mean : 7.155 Mean : 7.049

## 3rd Qu.: 9.000 3rd Qu.: 9.000 3rd Qu.: 8.000 3rd Qu.: 8.000

## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000

## NA's :6362 NA's :6362 NA's :6362 NA's :6362

str(dating)

## 'data.frame': 8378 obs. of 195 variables:

## $ iid : int 1 1 1 1 1 1 1 1 1 1 ...

## $ id : int 1 1 1 1 1 1 1 1 1 1 ...

## $ gender : int 0 0 0 0 0 0 0 0 0 0 ...

## $ idg : int 1 1 1 1 1 1 1 1 1 1 ...

## $ condtn : int 1 1 1 1 1 1 1 1 1 1 ...

## $ wave : int 1 1 1 1 1 1 1 1 1 1 ...

## $ round : int 10 10 10 10 10 10 10 10 10 10 ...

## $ position: int 7 7 7 7 7 7 7 7 7 7 ...

## $ positin1: int NA NA NA NA NA NA NA NA NA NA ...

## $ order : int 4 3 10 5 7 6 1 2 8 9 ...

## $ partner : int 1 2 3 4 5 6 7 8 9 10 ...

## $ pid : int 11 12 13 14 15 16 17 18 19 20 ...

## $ match : int 0 0 1 1 1 0 0 0 1 0 ...

## $ int\_corr: num 0.14 0.54 0.16 0.61 0.21 0.25 0.34 0.5 0.28 -0.36 ...

## $ samerace: int 0 0 1 0 0 0 0 0 0 0 ...

## $ age\_o : int 27 22 22 23 24 25 30 27 28 24 ...

## $ race\_o : int 2 2 4 2 3 2 2 2 2 2 ...

## $ pf\_o\_att: num 35 60 19 30 30 ...

## $ pf\_o\_sin: num 20 0 18 5 10 ...

## $ pf\_o\_int: num 20 0 19 15 20 ...

## $ pf\_o\_fun: num 20 40 18 40 10 ...

## $ pf\_o\_amb: num 0 0 14 5 10 ...

## $ pf\_o\_sha: num 5 0 12 5 20 ...

## $ dec\_o : int 0 0 1 1 1 1 0 0 1 0 ...

## $ attr\_o : num 6 7 10 7 8 7 3 6 7 6 ...

## $ sinc\_o : num 8 8 10 8 7 7 6 7 7 6 ...

## $ intel\_o : num 8 10 10 9 9 8 7 5 8 6 ...

## $ fun\_o : num 8 7 10 8 6 8 5 6 8 6 ...

## $ amb\_o : num 8 7 10 9 9 7 8 8 8 6 ...

## $ shar\_o : num 6 5 10 8 7 7 7 6 9 6 ...

## $ like\_o : num 7 8 10 7 8 7 2 7 6.5 6 ...

## $ prob\_o : num 4 4 10 7 6 6 1 5 8 6 ...

## $ met\_o : int 2 2 1 2 2 2 2 2 2 2 ...

## $ age : int 21 21 21 21 21 21 21 21 21 21 ...

## $ field : Factor w/ 260 levels "","Acting","African-American Studies/History",..: 152 152 152 152 152 152 152 152 152 152 ...

## $ field\_cd: num 1 1 1 1 1 1 1 1 1 1 ...

## $ undergra: Factor w/ 242 levels "","American University",..: 1 1 1 1 1 1 1 1 1 1 ...

## $ mn\_sat : Factor w/ 69 levels "","1,011.00",..: 1 1 1 1 1 1 1 1 1 1 ...

## $ tuition : Factor w/ 116 levels "","10,052.00",..: 1 1 1 1 1 1 1 1 1 1 ...

## $ race : int 4 4 4 4 4 4 4 4 4 4 ...

## $ imprace : int 2 2 2 2 2 2 2 2 2 2 ...

## $ imprelig: int 4 4 4 4 4 4 4 4 4 4 ...

## $ from : Factor w/ 270 levels "","94115","alabama",..: 56 56 56 56 56 56 56 56 56 56 ...

## $ zipcode : Factor w/ 410 levels "","0","1,040",..: 262 262 262 262 262 262 262 262 262 262 ...

## $ income : Factor w/ 262 levels "","106,663.00",..: 239 239 239 239 239 239 239 239 239 239 ...

## $ goal : int 2 2 2 2 2 2 2 2 2 2 ...

## $ date : int 7 7 7 7 7 7 7 7 7 7 ...

## $ go\_out : int 1 1 1 1 1 1 1 1 1 1 ...

## $ career : Factor w/ 368 levels "","?","??","a research position",..: 185 185 185 185 185 185 185 185 185 185 ...

## $ career\_c: num NA NA NA NA NA NA NA NA NA NA ...

## $ sports : int 9 9 9 9 9 9 9 9 9 9 ...

## $ tvsports: int 2 2 2 2 2 2 2 2 2 2 ...

## $ exercise: int 8 8 8 8 8 8 8 8 8 8 ...

## $ dining : int 9 9 9 9 9 9 9 9 9 9 ...

## $ museums : int 1 1 1 1 1 1 1 1 1 1 ...

## $ art : int 1 1 1 1 1 1 1 1 1 1 ...

## $ hiking : int 5 5 5 5 5 5 5 5 5 5 ...

## $ gaming : int 1 1 1 1 1 1 1 1 1 1 ...

## $ clubbing: int 5 5 5 5 5 5 5 5 5 5 ...

## $ reading : int 6 6 6 6 6 6 6 6 6 6 ...

## $ tv : int 9 9 9 9 9 9 9 9 9 9 ...

## $ theater : int 1 1 1 1 1 1 1 1 1 1 ...

## $ movies : int 10 10 10 10 10 10 10 10 10 10 ...

## $ concerts: int 10 10 10 10 10 10 10 10 10 10 ...

## $ music : int 9 9 9 9 9 9 9 9 9 9 ...

## $ shopping: int 8 8 8 8 8 8 8 8 8 8 ...

## $ yoga : int 1 1 1 1 1 1 1 1 1 1 ...

## $ exphappy: int 3 3 3 3 3 3 3 3 3 3 ...

## $ expnum : int 2 2 2 2 2 2 2 2 2 2 ...

## $ attr1\_1 : num 15 15 15 15 15 15 15 15 15 15 ...

## $ sinc1\_1 : num 20 20 20 20 20 20 20 20 20 20 ...

## $ intel1\_1: num 20 20 20 20 20 20 20 20 20 20 ...

## $ fun1\_1 : num 15 15 15 15 15 15 15 15 15 15 ...

## $ amb1\_1 : num 15 15 15 15 15 15 15 15 15 15 ...

## $ shar1\_1 : num 15 15 15 15 15 15 15 15 15 15 ...

## $ attr4\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ sinc4\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ intel4\_1: int NA NA NA NA NA NA NA NA NA NA ...

## $ fun4\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ amb4\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ shar4\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ attr2\_1 : num 35 35 35 35 35 35 35 35 35 35 ...

## $ sinc2\_1 : num 20 20 20 20 20 20 20 20 20 20 ...

## $ intel2\_1: num 15 15 15 15 15 15 15 15 15 15 ...

## $ fun2\_1 : num 20 20 20 20 20 20 20 20 20 20 ...

## $ amb2\_1 : num 5 5 5 5 5 5 5 5 5 5 ...

## $ shar2\_1 : num 5 5 5 5 5 5 5 5 5 5 ...

## $ attr3\_1 : int 6 6 6 6 6 6 6 6 6 6 ...

## $ sinc3\_1 : int 8 8 8 8 8 8 8 8 8 8 ...

## $ fun3\_1 : int 8 8 8 8 8 8 8 8 8 8 ...

## $ intel3\_1: int 8 8 8 8 8 8 8 8 8 8 ...

## $ amb3\_1 : int 7 7 7 7 7 7 7 7 7 7 ...

## $ attr5\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ sinc5\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ intel5\_1: int NA NA NA NA NA NA NA NA NA NA ...

## $ fun5\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ amb5\_1 : int NA NA NA NA NA NA NA NA NA NA ...

## $ dec : int 1 1 1 1 1 0 1 0 1 1 ...

## $ attr : num 6 7 5 7 5 4 7 4 7 5 ...

## [list output truncated]

head(dating)

## iid id gender idg condtn wave round position positin1 order partner pid

## 1 1 1 0 1 1 1 10 7 NA 4 1 11

## 2 1 1 0 1 1 1 10 7 NA 3 2 12

## 3 1 1 0 1 1 1 10 7 NA 10 3 13

## 4 1 1 0 1 1 1 10 7 NA 5 4 14

## 5 1 1 0 1 1 1 10 7 NA 7 5 15

## 6 1 1 0 1 1 1 10 7 NA 6 6 16

## match int\_corr samerace age\_o race\_o pf\_o\_att pf\_o\_sin pf\_o\_int pf\_o\_fun

## 1 0 0.14 0 27 2 35 20 20 20

## 2 0 0.54 0 22 2 60 0 0 40

## 3 1 0.16 1 22 4 19 18 19 18

## 4 1 0.61 0 23 2 30 5 15 40

## 5 1 0.21 0 24 3 30 10 20 10

## 6 0 0.25 0 25 2 50 0 30 10

## pf\_o\_amb pf\_o\_sha dec\_o attr\_o sinc\_o intel\_o fun\_o amb\_o shar\_o like\_o

## 1 0 5 0 6 8 8 8 8 6 7

## 2 0 0 0 7 8 10 7 7 5 8

## 3 14 12 1 10 10 10 10 10 10 10

## 4 5 5 1 7 8 9 8 9 8 7

## 5 10 20 1 8 7 9 6 9 7 8

## 6 0 10 1 7 7 8 8 7 7 7

## prob\_o met\_o age field field\_cd undergra mn\_sat tuition race imprace

## 1 4 2 21 Law 1 4 2

## 2 4 2 21 Law 1 4 2

## 3 10 1 21 Law 1 4 2

## 4 7 2 21 Law 1 4 2

## 5 6 2 21 Law 1 4 2

## 6 6 2 21 Law 1 4 2

## imprelig from zipcode income goal date go\_out career career\_c

## 1 4 Chicago 60,521 69,487.00 2 7 1 lawyer NA

## 2 4 Chicago 60,521 69,487.00 2 7 1 lawyer NA

## 3 4 Chicago 60,521 69,487.00 2 7 1 lawyer NA

## 4 4 Chicago 60,521 69,487.00 2 7 1 lawyer NA

## 5 4 Chicago 60,521 69,487.00 2 7 1 lawyer NA

## 6 4 Chicago 60,521 69,487.00 2 7 1 lawyer NA

## sports tvsports exercise dining museums art hiking gaming clubbing

## 1 9 2 8 9 1 1 5 1 5

## 2 9 2 8 9 1 1 5 1 5

## 3 9 2 8 9 1 1 5 1 5

## 4 9 2 8 9 1 1 5 1 5

## 5 9 2 8 9 1 1 5 1 5

## 6 9 2 8 9 1 1 5 1 5

## reading tv theater movies concerts music shopping yoga exphappy expnum

## 1 6 9 1 10 10 9 8 1 3 2

## 2 6 9 1 10 10 9 8 1 3 2

## 3 6 9 1 10 10 9 8 1 3 2

## 4 6 9 1 10 10 9 8 1 3 2

## 5 6 9 1 10 10 9 8 1 3 2

## 6 6 9 1 10 10 9 8 1 3 2

## attr1\_1 sinc1\_1 intel1\_1 fun1\_1 amb1\_1 shar1\_1 attr4\_1 sinc4\_1 intel4\_1

## 1 15 20 20 15 15 15 NA NA NA

## 2 15 20 20 15 15 15 NA NA NA

## 3 15 20 20 15 15 15 NA NA NA

## 4 15 20 20 15 15 15 NA NA NA

## 5 15 20 20 15 15 15 NA NA NA

## 6 15 20 20 15 15 15 NA NA NA

## fun4\_1 amb4\_1 shar4\_1 attr2\_1 sinc2\_1 intel2\_1 fun2\_1 amb2\_1 shar2\_1

## 1 NA NA NA 35 20 15 20 5 5

## 2 NA NA NA 35 20 15 20 5 5

## 3 NA NA NA 35 20 15 20 5 5

## 4 NA NA NA 35 20 15 20 5 5

## 5 NA NA NA 35 20 15 20 5 5

## 6 NA NA NA 35 20 15 20 5 5

## attr3\_1 sinc3\_1 fun3\_1 intel3\_1 amb3\_1 attr5\_1 sinc5\_1 intel5\_1 fun5\_1

## 1 6 8 8 8 7 NA NA NA NA

## 2 6 8 8 8 7 NA NA NA NA

## 3 6 8 8 8 7 NA NA NA NA

## 4 6 8 8 8 7 NA NA NA NA

## 5 6 8 8 8 7 NA NA NA NA

## 6 6 8 8 8 7 NA NA NA NA

## amb5\_1 dec attr sinc intel fun amb shar like prob met match\_es attr1\_s

## 1 NA 1 6 9 7 7 6 5 7 6 2 4 NA

## 2 NA 1 7 8 7 8 5 6 7 5 1 4 NA

## 3 NA 1 5 8 9 8 5 7 7 NA 1 4 NA

## 4 NA 1 7 6 8 7 6 8 7 6 2 4 NA

## 5 NA 1 5 6 7 7 6 6 6 6 2 4 NA

## 6 NA 0 4 9 7 4 6 4 6 5 2 4 NA

## sinc1\_s intel1\_s fun1\_s amb1\_s shar1\_s attr3\_s sinc3\_s intel3\_s fun3\_s

## 1 NA NA NA NA NA NA NA NA NA

## 2 NA NA NA NA NA NA NA NA NA

## 3 NA NA NA NA NA NA NA NA NA

## 4 NA NA NA NA NA NA NA NA NA

## 5 NA NA NA NA NA NA NA NA NA

## 6 NA NA NA NA NA NA NA NA NA

## amb3\_s satis\_2 length numdat\_2 attr7\_2 sinc7\_2 intel7\_2 fun7\_2 amb7\_2

## 1 NA 6 2 1 NA NA NA NA NA

## 2 NA 6 2 1 NA NA NA NA NA

## 3 NA 6 2 1 NA NA NA NA NA

## 4 NA 6 2 1 NA NA NA NA NA

## 5 NA 6 2 1 NA NA NA NA NA

## 6 NA 6 2 1 NA NA NA NA NA

## shar7\_2 attr1\_2 sinc1\_2 intel1\_2 fun1\_2 amb1\_2 shar1\_2 attr4\_2 sinc4\_2

## 1 NA 19.44 16.67 13.89 22.22 11.11 16.67 NA NA

## 2 NA 19.44 16.67 13.89 22.22 11.11 16.67 NA NA

## 3 NA 19.44 16.67 13.89 22.22 11.11 16.67 NA NA

## 4 NA 19.44 16.67 13.89 22.22 11.11 16.67 NA NA

## 5 NA 19.44 16.67 13.89 22.22 11.11 16.67 NA NA

## 6 NA 19.44 16.67 13.89 22.22 11.11 16.67 NA NA

## intel4\_2 fun4\_2 amb4\_2 shar4\_2 attr2\_2 sinc2\_2 intel2\_2 fun2\_2 amb2\_2

## 1 NA NA NA NA NA NA NA NA NA

## 2 NA NA NA NA NA NA NA NA NA

## 3 NA NA NA NA NA NA NA NA NA

## 4 NA NA NA NA NA NA NA NA NA

## 5 NA NA NA NA NA NA NA NA NA

## 6 NA NA NA NA NA NA NA NA NA

## shar2\_2 attr3\_2 sinc3\_2 intel3\_2 fun3\_2 amb3\_2 attr5\_2 sinc5\_2 intel5\_2

## 1 NA 6 7 8 7 6 NA NA NA

## 2 NA 6 7 8 7 6 NA NA NA

## 3 NA 6 7 8 7 6 NA NA NA

## 4 NA 6 7 8 7 6 NA NA NA

## 5 NA 6 7 8 7 6 NA NA NA

## 6 NA 6 7 8 7 6 NA NA NA

## fun5\_2 amb5\_2 you\_call them\_cal date\_3 numdat\_3 num\_in\_3 attr1\_3 sinc1\_3

## 1 NA NA 1 1 0 NA NA 15 20

## 2 NA NA 1 1 0 NA NA 15 20

## 3 NA NA 1 1 0 NA NA 15 20

## 4 NA NA 1 1 0 NA NA 15 20

## 5 NA NA 1 1 0 NA NA 15 20

## 6 NA NA 1 1 0 NA NA 15 20

## intel1\_3 fun1\_3 amb1\_3 shar1\_3 attr7\_3 sinc7\_3 intel7\_3 fun7\_3 amb7\_3

## 1 20 15 15 15 NA NA NA NA NA

## 2 20 15 15 15 NA NA NA NA NA

## 3 20 15 15 15 NA NA NA NA NA

## 4 20 15 15 15 NA NA NA NA NA

## 5 20 15 15 15 NA NA NA NA NA

## 6 20 15 15 15 NA NA NA NA NA

## shar7\_3 attr4\_3 sinc4\_3 intel4\_3 fun4\_3 amb4\_3 shar4\_3 attr2\_3 sinc2\_3

## 1 NA NA NA NA NA NA NA NA NA

## 2 NA NA NA NA NA NA NA NA NA

## 3 NA NA NA NA NA NA NA NA NA

## 4 NA NA NA NA NA NA NA NA NA

## 5 NA NA NA NA NA NA NA NA NA

## 6 NA NA NA NA NA NA NA NA NA

## intel2\_3 fun2\_3 amb2\_3 shar2\_3 attr3\_3 sinc3\_3 intel3\_3 fun3\_3 amb3\_3

## 1 NA NA NA NA 5 7 7 7 7

## 2 NA NA NA NA 5 7 7 7 7

## 3 NA NA NA NA 5 7 7 7 7

## 4 NA NA NA NA 5 7 7 7 7

## 5 NA NA NA NA 5 7 7 7 7

## 6 NA NA NA NA 5 7 7 7 7

## attr5\_3 sinc5\_3 intel5\_3 fun5\_3 amb5\_3

## 1 NA NA NA NA NA

## 2 NA NA NA NA NA

## 3 NA NA NA NA NA

## 4 NA NA NA NA NA

## 5 NA NA NA NA NA

## 6 NA NA NA NA NA

tail(dating)

## iid id gender idg condtn wave round position positin1 order partner

## 8373 552 22 1 44 2 21 22 17 10 8 17

## 8374 552 22 1 44 2 21 22 14 10 5 18

## 8375 552 22 1 44 2 21 22 13 10 4 19

## 8376 552 22 1 44 2 21 22 19 10 10 20

## 8377 552 22 1 44 2 21 22 3 10 16 21

## 8378 552 NA 1 44 2 21 22 2 10 15 22

## pid match int\_corr samerace age\_o race\_o pf\_o\_att pf\_o\_sin pf\_o\_int

## 8373 525 0 0.28 1 24 2 10 15 30

## 8374 526 0 0.64 0 26 3 10 10 30

## 8375 527 0 0.71 0 24 6 50 20 10

## 8376 528 0 -0.46 0 29 3 40 10 30

## 8377 529 0 0.62 0 22 4 10 25 25

## 8378 530 0 0.01 0 22 4 20 20 10

## pf\_o\_fun pf\_o\_amb pf\_o\_sha dec\_o attr\_o sinc\_o intel\_o fun\_o amb\_o

## 8373 20 15 10 0 8 8 7 7 8

## 8374 20 10 15 1 10 5 3 2 6

## 8375 5 10 5 0 6 3 7 3 7

## 8376 10 10 NA 0 2 1 2 2 2

## 8377 10 10 20 1 5 7 5 5 3

## 8378 15 5 30 1 8 8 7 7 7

## shar\_o like\_o prob\_o met\_o age field field\_cd

## 8373 6 7 4 2 25 Climate Dynamics 18

## 8374 5 6 1 NA 25 Climate Dynamics 18

## 8375 2 2 2 2 25 Climate Dynamics 18

## 8376 1 2 1 2 25 Climate Dynamics 18

## 8377 6 6 4 2 25 Climate Dynamics 18

## 8378 7 8 5 2 25 Climate Dynamics 18

## undergra mn\_sat tuition race imprace imprelig

## 8373 Ecole Normale Suprieure, Paris 2 1 1

## 8374 Ecole Normale Suprieure, Paris 2 1 1

## 8375 Ecole Normale Suprieure, Paris 2 1 1

## 8376 Ecole Normale Suprieure, Paris 2 1 1

## 8377 Ecole Normale Suprieure, Paris 2 1 1

## 8378 Ecole Normale Suprieure, Paris 2 1 1

## from zipcode income goal date go\_out

## 8373 France 78,110 1 2 1

## 8374 France 78,110 1 2 1

## 8375 France 78,110 1 2 1

## 8376 France 78,110 1 2 1

## 8377 France 78,110 1 2 1

## 8378 France 78,110 1 2 1

## career

## 8373 assistant master of the universe (otherwise it's too much work)

## 8374 assistant master of the universe (otherwise it's too much work)

## 8375 assistant master of the universe (otherwise it's too much work)

## 8376 assistant master of the universe (otherwise it's too much work)

## 8377 assistant master of the universe (otherwise it's too much work)

## 8378 assistant master of the universe (otherwise it's too much work)

## career\_c sports tvsports exercise dining museums art hiking gaming

## 8373 15 8 2 5 10 10 10 7 1

## 8374 15 8 2 5 10 10 10 7 1

## 8375 15 8 2 5 10 10 10 7 1

## 8376 15 8 2 5 10 10 10 7 1

## 8377 15 8 2 5 10 10 10 7 1

## 8378 15 8 2 5 10 10 10 7 1

## clubbing reading tv theater movies concerts music shopping yoga

## 8373 9 8 3 7 9 10 10 7 3

## 8374 9 8 3 7 9 10 10 7 3

## 8375 9 8 3 7 9 10 10 7 3

## 8376 9 8 3 7 9 10 10 7 3

## 8377 9 8 3 7 9 10 10 7 3

## 8378 9 8 3 7 9 10 10 7 3

## exphappy expnum attr1\_1 sinc1\_1 intel1\_1 fun1\_1 amb1\_1 shar1\_1

## 8373 10 NA 70 0 15 15 0 0

## 8374 10 NA 70 0 15 15 0 0

## 8375 10 NA 70 0 15 15 0 0

## 8376 10 NA 70 0 15 15 0 0

## 8377 10 NA 70 0 15 15 0 0

## 8378 10 NA 70 0 15 15 0 0

## attr4\_1 sinc4\_1 intel4\_1 fun4\_1 amb4\_1 shar4\_1 attr2\_1 sinc2\_1

## 8373 90 0 0 0 0 10 50 0

## 8374 90 0 0 0 0 10 50 0

## 8375 90 0 0 0 0 10 50 0

## 8376 90 0 0 0 0 10 50 0

## 8377 90 0 0 0 0 10 50 0

## 8378 90 0 0 0 0 10 50 0

## intel2\_1 fun2\_1 amb2\_1 shar2\_1 attr3\_1 sinc3\_1 fun3\_1 intel3\_1 amb3\_1

## 8373 0 30 0 20 8 7 6 7 7

## 8374 0 30 0 20 8 7 6 7 7

## 8375 0 30 0 20 8 7 6 7 7

## 8376 0 30 0 20 8 7 6 7 7

## 8377 0 30 0 20 8 7 6 7 7

## 8378 0 30 0 20 8 7 6 7 7

## attr5\_1 sinc5\_1 intel5\_1 fun5\_1 amb5\_1 dec attr sinc intel fun amb

## 8373 9 7 10 5 9 0 7 5 5 5 6

## 8374 9 7 10 5 9 0 3 5 5 5 NA

## 8375 9 7 10 5 9 0 4 6 8 4 4

## 8376 9 7 10 5 9 0 4 7 8 8 8

## 8377 9 7 10 5 9 0 4 6 5 4 NA

## 8378 9 7 10 5 9 0 3 7 6 4 8

## shar like prob met match\_es attr1\_s sinc1\_s intel1\_s fun1\_s amb1\_s

## 8373 NA 4 4 0 3 NA NA NA NA NA

## 8374 NA 2 5 0 3 NA NA NA NA NA

## 8375 NA 4 4 0 3 NA NA NA NA NA

## 8376 NA 6 5 0 3 NA NA NA NA NA

## 8377 5 5 5 0 3 NA NA NA NA NA

## 8378 1 4 5 0 3 NA NA NA NA NA

## shar1\_s attr3\_s sinc3\_s intel3\_s fun3\_s amb3\_s satis\_2 length

## 8373 NA NA NA NA NA NA 5 1

## 8374 NA NA NA NA NA NA 5 1

## 8375 NA NA NA NA NA NA 5 1

## 8376 NA NA NA NA NA NA 5 1

## 8377 NA NA NA NA NA NA 5 1

## 8378 NA NA NA NA NA NA 5 1

## numdat\_2 attr7\_2 sinc7\_2 intel7\_2 fun7\_2 amb7\_2 shar7\_2 attr1\_2

## 8373 2 70 0 15 10 0 5 70

## 8374 2 70 0 15 10 0 5 70

## 8375 2 70 0 15 10 0 5 70

## 8376 2 70 0 15 10 0 5 70

## 8377 2 70 0 15 10 0 5 70

## 8378 2 70 0 15 10 0 5 70

## sinc1\_2 intel1\_2 fun1\_2 amb1\_2 shar1\_2 attr4\_2 sinc4\_2 intel4\_2

## 8373 0 15 10 0 5 80 0 5

## 8374 0 15 10 0 5 80 0 5

## 8375 0 15 10 0 5 80 0 5

## 8376 0 15 10 0 5 80 0 5

## 8377 0 15 10 0 5 80 0 5

## 8378 0 15 10 0 5 80 0 5

## fun4\_2 amb4\_2 shar4\_2 attr2\_2 sinc2\_2 intel2\_2 fun2\_2 amb2\_2 shar2\_2

## 8373 5 0 10 50 5 10 20 5 10

## 8374 5 0 10 50 5 10 20 5 10

## 8375 5 0 10 50 5 10 20 5 10

## 8376 5 0 10 50 5 10 20 5 10

## 8377 5 0 10 50 5 10 20 5 10

## 8378 5 0 10 50 5 10 20 5 10

## attr3\_2 sinc3\_2 intel3\_2 fun3\_2 amb3\_2 attr5\_2 sinc5\_2 intel5\_2

## 8373 9 3 7 6 9 9 3 9

## 8374 9 3 7 6 9 9 3 9

## 8375 9 3 7 6 9 9 3 9

## 8376 9 3 7 6 9 9 3 9

## 8377 9 3 7 6 9 9 3 9

## 8378 9 3 7 6 9 9 3 9

## fun5\_2 amb5\_2 you\_call them\_cal date\_3 numdat\_3 num\_in\_3 attr1\_3

## 8373 4 7 2 0 0 NA 1 70

## 8374 4 7 2 0 0 NA 1 70

## 8375 4 7 2 0 0 NA 1 70

## 8376 4 7 2 0 0 NA 1 70

## 8377 4 7 2 0 0 NA 1 70

## 8378 4 7 2 0 0 NA 1 70

## sinc1\_3 intel1\_3 fun1\_3 amb1\_3 shar1\_3 attr7\_3 sinc7\_3 intel7\_3

## 8373 0 20 10 0 0 70 0 20

## 8374 0 20 10 0 0 70 0 20

## 8375 0 20 10 0 0 70 0 20

## 8376 0 20 10 0 0 70 0 20

## 8377 0 20 10 0 0 70 0 20

## 8378 0 20 10 0 0 70 0 20

## fun7\_3 amb7\_3 shar7\_3 attr4\_3 sinc4\_3 intel4\_3 fun4\_3 amb4\_3 shar4\_3

## 8373 10 0 0 80 0 10 0 0 10

## 8374 10 0 0 80 0 10 0 0 10

## 8375 10 0 0 80 0 10 0 0 10

## 8376 10 0 0 80 0 10 0 0 10

## 8377 10 0 0 80 0 10 0 0 10

## 8378 10 0 0 80 0 10 0 0 10

## attr2\_3 sinc2\_3 intel2\_3 fun2\_3 amb2\_3 shar2\_3 attr3\_3 sinc3\_3

## 8373 50 5 10 20 10 5 8 5

## 8374 50 5 10 20 10 5 8 5

## 8375 50 5 10 20 10 5 8 5

## 8376 50 5 10 20 10 5 8 5

## 8377 50 5 10 20 10 5 8 5

## 8378 50 5 10 20 10 5 8 5

## intel3\_3 fun3\_3 amb3\_3 attr5\_3 sinc5\_3 intel5\_3 fun5\_3 amb5\_3

## 8373 7 6 7 9 5 9 5 6

## 8374 7 6 7 9 5 9 5 6

## 8375 7 6 7 9 5 9 5 6

## 8376 7 6 7 9 5 9 5 6

## 8377 7 6 7 9 5 9 5 6

## 8378 7 6 7 9 5 9 5 6

#and I do not have a super computer so this will need to be simplified to work with the resourses available.

dating <- dating[,-(1:2),drop=FALSE]

dating <- subset( dating, select = -c( idg : pid ))

dating$samerace <- NULL

dating <- subset( dating, select = -c( race\_o : met\_o ))

dating <- subset( dating, select = -c( tuition : from ))

dating <- subset( dating, select = -c( career : yoga ))

dating <- subset( dating, select = -c( attr1\_1 : amb5\_3 ))

dating$income <- NULL

dating$field <- NULL

#Rename the columns for clarity

colnames(dating) <- c('gender', 'matchYN', 'intRate', 'agePart', 'age', 'field', 'almaMater', 'brains', 'home',

'goal', 'dateFreq', 'socialScore', 'attitude', 'swagLevel')

Tidy1Dating<- dating

#Examine the dataset

summary(dating)

## gender matchYN intRate agePart

## Min. :0.0000 Min. :0.0000 Min. :-0.830 Min. :18.00

## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:-0.020 1st Qu.:24.00

## Median :1.0000 Median :0.0000 Median : 0.210 Median :26.00

## Mean :0.5006 Mean :0.1647 Mean : 0.196 Mean :26.36

## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.: 0.430 3rd Qu.:28.00

## Max. :1.0000 Max. :1.0000 Max. : 0.910 Max. :55.00

## NA's :158 NA's :104

## age field almaMater brains

## Min. :18.00 Min. : 1.000 :3464 :5245

## 1st Qu.:24.00 1st Qu.: 5.000 UC Berkeley: 107 1,400.00: 403

## Median :26.00 Median : 8.000 Harvard : 104 1,430.00: 262

## Mean :26.36 Mean : 7.662 Columbia : 95 1,290.00: 190

## 3rd Qu.:28.00 3rd Qu.:10.000 Yale : 86 1,450.00: 163

## Max. :55.00 Max. :18.000 NYU : 78 1,340.00: 146

## NA's :95 NA's :82 (Other) :4444 (Other) :1969

## home goal dateFreq socialScore

## :1064 Min. :1.000 Min. :1.000 Min. :1.000

## 0 : 355 1st Qu.:1.000 1st Qu.:4.000 1st Qu.:1.000

## 10,021 : 139 Median :2.000 Median :5.000 Median :2.000

## 10,027 : 128 Mean :2.122 Mean :5.007 Mean :2.158

## 10,025 : 121 3rd Qu.:2.000 3rd Qu.:6.000 3rd Qu.:3.000

## 19,087 : 48 Max. :6.000 Max. :7.000 Max. :7.000

## (Other):6523 NA's :79 NA's :97 NA's :79

## attitude swagLevel

## Min. : 1.000 Min. : 0.000

## 1st Qu.: 5.000 1st Qu.: 2.000

## Median : 6.000 Median : 4.000

## Mean : 5.534 Mean : 5.571

## 3rd Qu.: 7.000 3rd Qu.: 8.000

## Max. :10.000 Max. :20.000

## NA's :101 NA's :6578

str(dating)

## 'data.frame': 8378 obs. of 14 variables:

## $ gender : int 0 0 0 0 0 0 0 0 0 0 ...

## $ matchYN : int 0 0 1 1 1 0 0 0 1 0 ...

## $ intRate : num 0.14 0.54 0.16 0.61 0.21 0.25 0.34 0.5 0.28 -0.36 ...

## $ agePart : int 27 22 22 23 24 25 30 27 28 24 ...

## $ age : int 21 21 21 21 21 21 21 21 21 21 ...

## $ field : num 1 1 1 1 1 1 1 1 1 1 ...

## $ almaMater : Factor w/ 242 levels "","American University",..: 1 1 1 1 1 1 1 1 1 1 ...

## $ brains : Factor w/ 69 levels "","1,011.00",..: 1 1 1 1 1 1 1 1 1 1 ...

## $ home : Factor w/ 410 levels "","0","1,040",..: 262 262 262 262 262 262 262 262 262 262 ...

## $ goal : int 2 2 2 2 2 2 2 2 2 2 ...

## $ dateFreq : int 7 7 7 7 7 7 7 7 7 7 ...

## $ socialScore: int 1 1 1 1 1 1 1 1 1 1 ...

## $ attitude : int 3 3 3 3 3 3 3 3 3 3 ...

## $ swagLevel : int 2 2 2 2 2 2 2 2 2 2 ...

#swagLevel has too many NA's so I am going to remove that

dating$swagLevel <- NULL

#look closer at remaining columns

summary(dating$almaMater)

##

## 3464

## UC Berkeley

## 107

## Harvard

## 104

## Columbia

## 95

## Yale

## 86

## NYU

## 78

## Brown

## 66

## UCLA

## 66

## Cornell University

## 64

## Smith College

## 59

## Tufts University

## 47

## Columbia College

## 44

## Cornell

## 44

## Columbia University

## 43

## Rutgers College

## 43

## University of Toronto

## 43

## Barnard College

## 42

## University of Pennsylvania

## 41

## Wesleyan University

## 40

## Harvard University

## 35

## University of Michigan

## 34

## Penn State University

## 33

## University of Vermont

## 33

## Connecticut College

## 31

## Brown University

## 30

## Florida International University

## 30

## Tufts

## 30

## University of Washington

## 30

## Delhi University

## 29

## Georgetown University

## 28

## Holy Cross

## 28

## Princeton

## 28

## UC Irvine

## 28

## University of California at Santa Barbara

## 28

## Boston College

## 27

## Stanford University

## 25

## Amherst College

## 24

## Fordham University

## 24

## Bucknell University

## 23

## Rutgers University

## 23

## Beijing University

## 22

## California State University Los Angeles

## 22

## ColumbiaU

## 22

## Conneticut College

## 22

## COOPER UNION

## 22

## Cooper Union, Bard college, and SUNY Purchase

## 22

## Dartmouth College

## 22

## Ecole Normale Suprieure, Paris

## 22

## Georgetown

## 22

## Hamilton College

## 22

## Loyola College

## 22

## LUISS, Rome

## 22

## Naples, Italy

## 22

## Nirma Institute of Technology-India

## 22

## Oxford

## 22

## Oxford University

## 22

## Princeton University

## 22

## S.V Regional Engineering College,India

## 22

## Santa Clara University

## 22

## SEAS

## 22

## Southwestern University

## 22

## Texas State University

## 22

## Tianjin University in China

## 22

## U.C. Berkeley

## 22

## Univeristy of California, Davis

## 22

## Universidad de Chile

## 22

## University of Cologne, Germany

## 22

## University of Genova

## 22

## University of Kansas

## 22

## University of Reading, England

## 22

## University of Southern California

## 22

## Brandeis University

## 21

## charles university, prague, czech republic

## 21

## ChungShenMedicalUniversity(Taiwan)

## 21

## colby college, waterville, me

## 21

## Columbia Business School

## 21

## Emory University

## 21

## Harvey Mudd College (Physics)

## 21

## MIT

## 21

## MSU, Russia

## 21

## Novosibirsk State University

## 21

## PACE University

## 21

## Queens College

## 21

## REC, Rourkela

## 21

## Rice

## 21

## RPI

## 21

## Sarah Lawrence College

## 21

## Stanford

## 21

## Syracuse University

## 21

## tech school

## 21

## Tokyo Woman's Christian University, Japan

## 21

## U of Michigan

## 21

## umass

## 21

## Universidad Iberoamericana

## 21

## University of Maryland, and Oxford

## 21

## University of North Carolina at Charlotte

## 21

## University of Paris

## 21

## University of Warsaw

## 21

## UNLV

## 21

## (Other)

## 1931

#the number of factors needs to be reduced time to create new catagories.

dating$almaMater <- gsub('^University of Toronto', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^Universidad Iberoamericana', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^Universidad de Chile', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^Universidad de Costa Rica', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^Rutgers University - New Brunswick', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^school of social sciences in uruguay', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^U. del Rosario, Medicine, Colombia SA', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^Universidad Catolica de Chile', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^Universidad de los Andes', 'Foreign American School', dating$almaMater)

dating$almaMater <- gsub('^McGill University', 'Foreign American School', dating$almaMater)

#(this is also where I learned that you need to put Oxford University before Oxford, for example, or you end up with almaMater University)

dating$almaMater <- gsub('^LUISS, Rome', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Oxford University', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Oxford', 'European School', dating$almaMater)

dating$almaMater <- gsub('^University of Cologne, Germany', 'European School', dating$almaMater)

dating$almaMater <- gsub('^University of Reading, England', 'European School', dating$almaMater)

dating$almaMater <- gsub('^charles university, prague, czech republic', 'European School', dating$almaMater)

dating$almaMater <- gsub('^University of Paris', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Naples, Italy', 'European School', dating$almaMater)

dating$almaMater <- gsub('^University of Genova', 'European School', dating$almaMater)

dating$almaMater <- gsub('^University of Warsaw', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Ecole Normale SupZrieure, Paris', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Novosibirsk State University', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Univeristy of Barcelona', 'European School', dating$almaMater)

##dating$almaMater <- gsub('^European School University', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Cambridge University', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Bocconi University Milan', 'European School', dating$almaMater)

dating$almaMater <- gsub('^ecole polytechnique', 'European School', dating$almaMater)

dating$almaMater <- gsub('^HEC FRance', 'European School', dating$almaMater)

dating$almaMater <- gsub('^MSU, Russia', 'European School', dating$almaMater)

##dating$almaMater <- gsub('^Supaero (France)', 'European School', dating$almaMater) #NOT WORKING!!!!!!!!!!!!!!!!!!!!!!

dating$almaMater <- gsub('^University of Heidelberg', 'European School', dating$almaMater)

dating$almaMater <- gsub('^University of Karlsruhe/Germany', 'European School', dating$almaMater)

dating$almaMater <- gsub('^warsaw university', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Bucharest University', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Cagliari - Italy', 'European School', dating$almaMater)

dating$almaMater <- gsub('^Ecole Polytechnique (France)', 'European School', dating$almaMater)

##dating$almaMater <- gsub('^Ecole Superieure d'Electricite', 'European School', dating$almaMater) !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

dating$almaMater <- gsub('^S.V Regional Engineering College,India', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Delhi University', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Beijing University', 'Asian School', dating$almaMater)

###dating$almaMater <- gsub('^ChungShenMedicalUniversity(Taiwan)', 'Asian School', dating$almaMater) ##NOT WORKING!!!!!!!!!!!!!!!!!!!!!

dating$almaMater <- gsub('^MSU, Russia(Taiwan)', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Nirma Institute of Technology-India', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Tianjin University in China', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^REC, Rourkela', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Monash University - Australia', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^SOGANG UNIVERSITY(KOREA)', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Fudan University, Shanghai, China', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Fudan', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Harbin Medical University, China', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Kunitachi College of Music (in Japan)', 'Asian School', dating$almaMater) ##NOT WORKING!!!!!!!!!!!!!!!!!!!

dating$almaMater <- gsub('^National University of Singapore', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Rizvi College of Architecture, Bombay University', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Shia-Chian University', 'Asian School', dating$almaMater)

###dating$almaMater <- gsub('^SOGANG UNIVERSITY(KOREA)', 'Asian School', dating$almaMater) ##NOT WORKING!!!!!!!!!!!!!!!!!!!

dating$almaMater <- gsub('^Univ. of Bombay', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^University of Delhi', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^University of the Philippines', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Ateneo de Manila University - Philippines', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Bombay, India', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Fu Jen Catholic University, Taiwan', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Jesus and Mary College,Delhi', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^National Taiwan University', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^National University Of Singapore', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Taiwan University', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^university of the philippines', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^university of the philippines', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^China', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Harcourt Butler Technological Institute', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^Hebrew University', 'Asian School', dating$almaMater)

dating$almaMater <- gsub('^NUS', 'Asian School', dating$almaMater)

#dating$almaMater <- gsub('^Tokyo Woman's Christian University, Japan', 'Asian School', dating$almaMater) NOTWORKING!!!!!!!!!!!!!!!!!!!!!

dating$almaMater <- gsub('^Columbia Business School', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Columbia College, CU', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Columbia College', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^ColumbiaU', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Columbia', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Brown University', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Brown', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Cornell University', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Cornell', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Dartmouth College', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Harvard University', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Harvard College', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Harvard', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^harvard', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Princeton University', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Princeton U.', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Princeton', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^University of Pennsylvania', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^university of pennsylvania', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Univ of Pennsylvania', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^UPenn', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Yale University', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Yale', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('^Ivy University', 'Ivy', dating$almaMater)

dating$almaMater <- gsub('Amherst College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Bowdoin College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Bucknell University', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Colgate University', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Conneticut College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Connecticut College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('colby college, waterville, me', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Hamilton College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Lafayette College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Middlebury College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Swarthmore College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Trinity College', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Tufts University', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Tufts', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('tufts', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('Wesleyan University', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('LittleIvy University', 'LittleIvy', dating$almaMater)

dating$almaMater <- gsub('^wellesley college', 'Ladies', dating$almaMater)

dating$almaMater <- gsub('^Barnard College', 'Ladies', dating$almaMater)

dating$almaMater <- gsub('^Smith College', 'Ladies', dating$almaMater)

dating$almaMater <- gsub('^wellesley college', 'Ladies', dating$almaMater)

dating$almaMater <- gsub('Arizona State', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Colorado State', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Florida International University', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('GA Tech', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Miami University', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Rutgers College', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Rutgers University', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('SUNY Binghamton', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('SUNY Stony Brook', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('SUNY Albany', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('SUNY Geneseo', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Texas A&M', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Binghamton University', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('U of Michigan', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('U of Vermont', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Univ of New Mexico', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Univ. of Connecticut', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Akron, OH', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Arizona', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Florida', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Delaware', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Illinois/Champaign', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Kansas', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Maryland, and Oxford', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Massachusetts-Amherst', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Michigan', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Michigan-Ann Arbor', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of North Carolina at Charlotte', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Oregon', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Tennessee', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Texas', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Vermont', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Washington', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('University of Wisconsin-Madison', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('university of wisconsin/la crosse', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('UW Madison', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('UNLV', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Penn State University', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('PublicSchool-Ann Arbor', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Illinois', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('UM', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('umass', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Purdue', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Univeristy of Michigan', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('George Mason University', 'PublicSchool', dating$almaMater)

dating$almaMater <- gsub('Berklee College Of Music', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('Cal Berkeley', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('Cal State Univ., Long Beach', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('California State University Los Angeles', 'CaliSchool', dating$almaMater)

##dating$almaMater <- gsub('Saint Mary's College of California', 'CaliSchool', dating$almaMater) !!!!!!!!!!!!!

dating$almaMater <- gsub('Santa Clara University', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('Stanford University', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('Stanford', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('u of southern california, economics', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('U.C. Berkeley', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('UC Berkeley', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('UC Davis', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('UC Irvine', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('UC Santa Cruz', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('UC, IRVINE!!!!!!!!!', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('ucla', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('UCLA', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('UCSB', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('Univeristy of California, Davis', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('University of California at Santa Barbara', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('University of California at Santa Cruz', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('University of Southern California', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('CSUN', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('Harvey Mudd College (Physics)', 'CaliSchool', dating$almaMater)

dating$almaMater <- gsub('Boston College', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('MIT', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('New York University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Notre Dame', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Northwestern University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('NYU', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('nyu', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Syracuse University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Texas State University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Vanderbilt University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('George Washington University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Fordham University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Georgetown University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Georgetown', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('GW', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Emory University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('American University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('HOWARD UNIVERSITY', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('John Hopkins', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('PACE University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Queens College', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('University of Chicago', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('University of Rochester', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Wake Forest', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Washington U. in St. Louis', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('washington university in st louis', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('Washington University in St. Louis', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('William and Mary', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('BigPrivate University', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('The BigPrivate', 'BigPrivate', dating$almaMater)

dating$almaMater <- gsub('^Loyola College in Maryland', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Loyola College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Bennington College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Brandeis University', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Case Western Reserve University', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^COOPER UNION', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Cooper Union, Bard college, and SUNY Purchase', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Duquesne University', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Augustana College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Hampshire College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Holy Cross', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^lipscomb university', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Mary Baldwin College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^marymount manhattan college', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Oberlin College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^oberlin', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Occidental College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Rice University', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Rice', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^RPI', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Sarah Lawrence College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Southwestern University', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Kettering University / GMI', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Southwestern University', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^SmallPrivate College', 'SmallPrivate', dating$almaMater)

dating$almaMater <- gsub('^Engineering', 'Other', dating$almaMater)

dating$almaMater <- gsub('^medicine', 'Other', dating$almaMater)

##dating$almaMater <- gsub('^school overseas (need a name ?)', 'Other', dating$almaMater) ###NOT WORKING!!!!!!!!!!!!!!!!!!

dating$almaMater <- gsub('^tech school', 'Other', dating$almaMater)

dating$almaMater <- gsub('^University of International Business & Economics', 'Other', dating$almaMater)

dating$almaMater <- gsub('^Biological Sciences', 'Other', dating$almaMater)

dating$almaMater <- gsub('^School of the Arts', 'Other', dating$almaMater)

dating$almaMater <- gsub('^SEAS', 'Other', dating$almaMater)

dating$almaMater <- as.factor(dating$almaMater)

levels(dating$almaMater)

## [1] ""

## [2] "Asian School"

## [3] "BigPrivate"

## [4] "CaliSchool"

## [5] "ChungShenMedicalUniversity(Taiwan)"

## [6] "Ecole Normale Suprieure, Paris"

## [7] "Ecole Polytechnique (France)"

## [8] "Ecole Superieure d'Electricite"

## [9] "European School"

## [10] "Foreign American School"

## [11] "Harvey Mudd College (Physics)"

## [12] "Ivy"

## [13] "Kunitachi College of Music (in Japan)"

## [14] "Ladies"

## [15] "LittleIvy"

## [16] "Other"

## [17] "PublicSchool"

## [18] "Saint Mary's College of California"

## [19] "school overseas (need a name ?)"

## [20] "SmallPrivate"

## [21] "SOGANG UNIVERSITY(KOREA)"

## [22] "Supaero (France)"

## [23] "Tokyo Woman's Christian University, Japan"

## [24] "Yeshiva University"

levels(dating$almaMater)[1] <- "Other"

levels(dating$almaMater)[5] <- "Asian School"

levels(dating$almaMater)[6] <- "European School"

levels(dating$almaMater)[7] <- "CaliSchool"

levels(dating$almaMater)[8]<- "Asian School"

levels(dating$almaMater)[14] <- "Asian School"

levels(dating$almaMater)[15] <- "Asian School"

levels(dating$almaMater)[11] <- "CaliSchool"

levels(dating$almaMater)[11] <- "Other"

levels(dating$almaMater)[12] <- "European School"

levels(dating$almaMater)[12] <- "SmallPrivate"

summary(dating$almaMater)

## Other

## 4347

## Asian School

## 472

## BigPrivate

## 544

## CaliSchool

## 911

## Ecole Normale Suprieure, Paris

## 22

## European School

## 835

## Foreign American School

## 178

## Ivy

## 882

## Kunitachi College of Music (in Japan)

## 9

## Ladies

## 119

## Saint Mary's College of California

## 14

## SmallPrivate

## 18

## Tokyo Woman's Christian University, Japan

## 21

## Yeshiva University

## 6

TidywalmaMater <- dating

#Other is a little too large, but i like how well everything else is distributed

#Now to slim down the number of factors in the brains column

#Looks like about 3000 people responded so I am going to split them into an approximately

#even set of three levels, "low", "med", and "high" and While I'm at it I will rename " " "na."

summary(dating$brains)

## 1,011.00 1,014.00 1,030.00 1,034.00 1,050.00 1,060.00 1,070.00

## 5245 19 18 40 14 21 18 9

## 1,080.00 1,090.00 1,092.00 1,097.00 1,099.00 1,100.00 1,105.00 1,110.00

## 6 29 14 14 9 21 19 30

## 1,130.00 1,134.00 1,140.00 1,149.00 1,155.00 1,157.00 1,159.00 1,160.00

## 43 7 18 21 30 18 20 42

## 1,178.00 1,180.00 1,185.00 1,188.00 1,191.00 1,200.00 1,206.00 1,210.00

## 10 18 14 80 8 34 33 61

## 1,212.00 1,214.00 1,215.00 1,220.00 1,227.00 1,230.00 1,239.00 1,242.00

## 14 9 9 21 22 6 22 36

## 1,250.00 1,258.00 1,260.00 1,267.00 1,270.00 1,280.00 1,290.00 1,308.00

## 53 42 133 6 10 22 190 21

## 1,309.00 1,310.00 1,320.00 1,330.00 1,331.00 1,340.00 1,341.00 1,360.00

## 138 80 42 84 18 146 18 116

## 1,365.00 1,370.00 1,380.00 1,400.00 1,402.00 1,410.00 1,430.00 1,450.00

## 19 19 61 403 24 27 262 163

## 1,460.00 1,470.00 1,490.00 914.00 990.00

## 100 10 21 8 20

levels(dating$brains)

## [1] "" "1,011.00" "1,014.00" "1,030.00" "1,034.00" "1,050.00"

## [7] "1,060.00" "1,070.00" "1,080.00" "1,090.00" "1,092.00" "1,097.00"

## [13] "1,099.00" "1,100.00" "1,105.00" "1,110.00" "1,130.00" "1,134.00"

## [19] "1,140.00" "1,149.00" "1,155.00" "1,157.00" "1,159.00" "1,160.00"

## [25] "1,178.00" "1,180.00" "1,185.00" "1,188.00" "1,191.00" "1,200.00"

## [31] "1,206.00" "1,210.00" "1,212.00" "1,214.00" "1,215.00" "1,220.00"

## [37] "1,227.00" "1,230.00" "1,239.00" "1,242.00" "1,250.00" "1,258.00"

## [43] "1,260.00" "1,267.00" "1,270.00" "1,280.00" "1,290.00" "1,308.00"

## [49] "1,309.00" "1,310.00" "1,320.00" "1,330.00" "1,331.00" "1,340.00"

## [55] "1,341.00" "1,360.00" "1,365.00" "1,370.00" "1,380.00" "1,400.00"

## [61] "1,402.00" "1,410.00" "1,430.00" "1,450.00" "1,460.00" "1,470.00"

## [67] "1,490.00" "914.00" "990.00"

levels(dating$brains)[1] <- "na"

levels(dating$brains)[2:43] <- "low"

levels(dating$brains)[3:18] <- "mid"

levels(dating$brains)[4:13] <- "high"

#Time to look at home

summary(dating$home)

## 0 10,021 10,027 10,025 19,087 7,410 92,064 10,012

## 1064 355 139 128 121 48 41 41 39

## 11,235 10,128 91,011 80,131 20,817 7,726 10,019 7,936 8,904

## 39 37 37 36 34 34 32 31 31

## 60,521 10,028 8,820 10,024 10,029 10,594 11,001 11,570 11,572

## 30 29 26 22 22 22 22 22 22

## 12,020 12,563 136 16,146 19,380 2,140 2,420 26,223 33,418

## 22 22 22 22 22 22 22 22 22

## 395,001 45,213 471,001 48,306 50,354 67,111 7,078 78,110 78,666

## 22 22 22 22 22 22 22 22 22

## 8,805 84,108 90,034 91,754 91,789 92,028 94,121 94,536 95,831

## 22 22 22 22 22 22 22 22 22

## 96,797 1,851 10,009 10,023 10,536 10,538 10,543 11,212 11,910

## 22 21 21 21 21 21 21 21 21

## 125,438 18,977 20,011 21,044 21,701 22,015 430,000 50,450 6,098

## 21 21 21 21 21 21 21 21 21

## 6,320 630,090 7,024 7,512 7,620 76,710 77,546 80,304 89,014

## 21 21 21 21 21 21 21 21 21

## 98,115 10,017 10,469 10,471 100,063 11,354 11,363 11,364 11,531

## 21 20 20 20 20 20 20 20 20

## 11,596 11,803 12,601 13,850 135,110 14,526 15,211 18,929 19,046

## 20 20 20 20 20 20 20 20 20

## (Other)

## 4351

levels (dating$home)

## [1] "" "0" "1,040" "1,114" "1,128"

## [6] "1,173" "1,720" "1,742" "1,801" "1,851"

## [11] "1,867" "1,890" "10,001" "10,002" "10,006"

## [16] "10,009" "10,012" "10,014" "10,016" "10,017"

## [21] "10,019" "10,021" "10,022" "10,023" "10,024"

## [26] "10,025" "10,027" "10,028" "10,029" "10,086"

## [31] "10,128" "10,301" "10,306" "10,454" "10,463"

## [36] "10,469" "10,471" "10,502" "10,514" "10,523"

## [41] "10,536" "10,538" "10,543" "10,594" "10,598"

## [46] "10,605" "10,706" "10,803" "10,804" "100"

## [51] "100,063" "108" "11,000" "11,001" "11,020"

## [56] "11,021" "11,023" "11,040" "11,104" "11,137"

## [61] "11,204" "11,212" "11,214" "11,215" "11,217"

## [66] "11,234" "11,235" "11,354" "11,363" "11,364"

## [71] "11,365" "11,373" "11,375" "11,419" "11,432"

## [76] "11,500" "11,531" "11,552" "11,561" "11,570"

## [81] "11,572" "11,576" "11,596" "11,733" "11,746"

## [86] "11,753" "11,754" "11,778" "11,793" "11,803"

## [91] "11,910" "110,003" "110,015" "110,060" "12,000"

## [96] "12,020" "12,302" "12,563" "12,590" "12,601"

## [101] "12,603" "125,438" "13,413" "13,827" "13,850"

## [106] "135,110" "136" "136,300" "14,043" "14,075"

## [111] "14,227" "14,526" "14,850" "15,146" "15,211"

## [116] "15,668" "16,146" "16,510" "16,803" "17,403"

## [121] "18,103" "18,603" "18,929" "18,977" "19,041"

## [126] "19,046" "19,087" "19,151" "19,335" "19,380"

## [131] "19,422" "19,454" "2,021" "2,115" "2,138"

## [136] "2,140" "2,155" "2,173" "2,420" "2,467"

## [141] "20,001" "20,011" "20,129" "20,782" "20,815"

## [146] "20,816" "20,817" "20,852" "20,853" "20,854"

## [151] "20,878" "20,903" "200,000" "200,065" "21,015"

## [156] "21,020" "21,044" "21,093" "21,209" "21,701"

## [161] "210,009" "22,003" "22,015" "22,066" "22,151"

## [166] "22,442" "23,060" "248,001" "26,223" "26,900"

## [171] "27,701" "27,870" "28,035" "28,387" "28,804"

## [176] "29,055" "29,501" "29,571" "3,031" "3,186"

## [181] "30,066" "30,071" "30,092" "30,100" "30,345"

## [186] "30,677" "300,151" "32,304" "32,780" "33,021"

## [191] "33,156" "33,183" "33,184" "33,414" "33,418"

## [196] "33,496" "33,511" "34,105" "35,223" "35,404"

## [201] "36,701" "37,204" "370,138" "38,119" "38,330"

## [206] "395,001" "4,605" "400,051" "411,101" "43,220"

## [211] "43,229" "430,000" "44,118" "44,147" "45,213"

## [216] "45,242" "45,243" "46,205" "46,815" "46,818"

## [221] "47,906" "47,920" "471,001" "48,070" "48,098"

## [226] "48,104" "48,124" "48,127" "48,302" "48,306"

## [231] "48,331" "48,334" "48,895" "5,401" "50,354"

## [236] "50,450" "519,000" "52,803" "53,012" "53,217"

## [241] "53,705" "54,449" "55,331" "55,345" "55,379"

## [246] "55,391" "55,424" "55,446" "560,032" "597,627"

## [251] "6,019" "6,098" "6,268" "6,320" "6,437"

## [256] "6,700" "6,878" "6,880" "6,883" "60,089"

## [261] "60,519" "60,521" "60,611" "62,150" "62,996"

## [266] "63,034" "63,131" "630,090" "64,129" "650,206"

## [271] "66,208" "66,610" "66,614" "67,111" "68,005"

## [276] "68,124" "7,024" "7,030" "7,032" "7,039"

## [281] "7,045" "7,050" "7,069" "7,076" "7,078"

## [286] "7,201" "7,304" "7,410" "7,512" "7,605"

## [291] "7,620" "7,624" "7,661" "7,675" "7,719"

## [296] "7,726" "7,733" "7,746" "7,747" "7,760"

## [301] "7,901" "7,936" "70,605" "703,007" "75,015"

## [306] "76,116" "76,513" "76,710" "77,024" "77,026"

## [311] "77,077" "77,095" "77,096" "77,401" "77,546"

## [316] "78,100" "78,110" "78,230" "78,666" "78,759"

## [321] "8,003" "8,025" "8,028" "8,034" "8,071"

## [326] "8,091" "8,225" "8,536" "8,540" "8,691"

## [331] "8,805" "8,816" "8,820" "8,854" "8,904"

## [336] "80,005" "80,110" "80,123" "80,131" "80,136"

## [341] "80,220" "80,304" "80,798" "807,931" "84,108"

## [346] "85,201" "85,283" "87,004" "87,110" "89,014"

## [351] "9,012" "9,971,200" "90,026" "90,031" "90,034"

## [356] "90,036" "90,210" "90,272" "90,291" "90,503"

## [361] "90,504" "90,620" "901" "91,011" "91,206"

## [366] "91,360" "91,730" "91,754" "91,789" "92,024"

## [371] "92,028" "92,037" "92,064" "92,075" "92,120"

## [376] "92,425" "92,602" "92,683" "92,821" "92,833"

## [381] "92,843" "92,869" "92,879" "93,108" "93,257"

## [386] "94,022" "94,043" "94,044" "94,108" "94,121"

## [391] "94,130" "94,133" "94,306" "94,403" "94,536"

## [396] "94,539" "94,596" "94,933" "94,941" "94,960"

## [401] "95,008" "95,060" "95,404" "95,695" "95,831"

## [406] "96,701" "96,797" "96,822" "98,115" "98,579"

#I would like to use this, but there is no way to simplify this really. An entire Zip code

#is a large enough area as it is. I am going to combine " " "0" and "(Other)" together at least.

levels(dating$home)[1:2] <- "Other"

dating$gender <- factor(dating$gender)

summary(dating$gender)

## 0 1

## 4184 4194

levels(dating$gender)[1]<- "female"

levels(dating$gender)[2]<- "male"

dating$matchYN <- factor(dating$matchYN)

summary(dating$matchYN)

## 0 1

## 6998 1380

levels(dating$matchYN)[1] <- "no"

levels(dating$matchYN)[2] <- "yes"

dating$field <- factor(dating$field)

levels(dating$field)[1] <- "Law"

levels(dating$field)[2] <- "Math"

levels(dating$field)[3] <- "Social Science"

levels(dating$field)[4] <- "Medical Science"

levels(dating$field)[5] <- "Engineering"

levels(dating$field)[6] <- "English"

levels(dating$field)[7] <- "History/Philosophy"

levels(dating$field)[8] <- "Business"

levels(dating$field)[9] <- "Education"

levels(dating$field)[10] <- "Science"

levels(dating$field)[11] <- "Social Work"

levels(dating$field)[12] <- "Undecided"

levels(dating$field)[13] <- "PolySci"

levels(dating$field)[14] <- "Film"

levels(dating$field)[15] <- "Fine Arts"

levels(dating$field)[16] <- "Languages"

levels(dating$field)[17] <- "Architecture"

levels(dating$field)[18] <- "Other"

summary(dating$field)

## Law Math Social Science

## 665 207 696

## Medical Science Engineering English

## 143 864 325

## History/Philosophy Business Education

## 241 1925 626

## Science Social Work Undecided

## 993 468 19

## PolySci Film Fine Arts

## 709 126 187

## Languages Architecture Other

## 40 10 52

## NA's

## 82

levels(dating$field)

## [1] "Law" "Math" "Social Science"

## [4] "Medical Science" "Engineering" "English"

## [7] "History/Philosophy" "Business" "Education"

## [10] "Science" "Social Work" "Undecided"

## [13] "PolySci" "Film" "Fine Arts"

## [16] "Languages" "Architecture" "Other"

#I could combine Film and Fine Arts maybe....gonna leave it alone for now

dating$goal <- factor(dating$goal)

levels(dating$goal)[1] <- "Seemed like a fun night out"

levels(dating$goal)[2] <- "To meet new people"

levels(dating$goal)[3] <- "To get a date"

levels(dating$goal)[4] <- "Looking for a serious relationship"

levels(dating$goal)[5] <- "To say I did it"

levels(dating$goal)[6] <- "Other"

summary(dating$goal)

## Seemed like a fun night out To meet new people

## 3426 3012

## To get a date Looking for a serious relationship

## 631 301

## To say I did it Other

## 510 419

## NA's

## 79

dating$dateFreq <- factor(dating$dateFreq)

levels(dating$dateFreq)[1] <- "Several times a week"

levels(dating$dateFreq)[2] <- "Twice a week"

levels(dating$dateFreq)[3] <- "Once a week"

levels(dating$dateFreq)[4] <- "Twice a month"

levels(dating$dateFreq)[5] <- "Once a month"

levels(dating$dateFreq)[6] <- "Several times a year"

levels(dating$dateFreq)[7] <- "Almost Never"

dating$socialScore <- factor(dating$socialScore)

levels(dating$socialScore)[1] <- "Several times a week"

levels(dating$socialScore)[2] <- "Twice a week"

levels(dating$socialScore)[3] <- "Once a week"

levels(dating$socialScore)[4] <- "Twice a month"

levels(dating$socialScore)[5] <- "Once a month"

levels(dating$socialScore)[6] <- "Several times a year"

levels(dating$socialScore)[7] <- "Almost never"

summary(dating$socialScore)

## Several times a week Twice a week Once a week

## 2610 2990 1949

## Twice a month Once a month Several times a year

## 450 164 99

## Almost never NA's

## 37 79

TidydataPrePurge <- dating

# create a dataframe

df1 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$almaMater))

names(df1) <- c('match', 'almaMater', 'count')

df1

## match almaMater count

## 1 no Other 3625

## 2 yes Other 722

## 3 no Asian School 410

## 4 yes Asian School 62

## 5 no BigPrivate 461

## 6 yes BigPrivate 83

## 7 no CaliSchool 782

## 8 yes CaliSchool 129

## 9 no Ecole Normale Suprieure, Paris 16

## 10 yes Ecole Normale Suprieure, Paris 6

## 11 no European School 680

## 12 yes European School 155

## 13 no Foreign American School 146

## 14 yes Foreign American School 32

## 15 no Ivy 726

## 16 yes Ivy 156

## 17 no Kunitachi College of Music (in Japan) 7

## 18 yes Kunitachi College of Music (in Japan) 2

## 19 no Ladies 95

## 20 yes Ladies 24

## 21 no Saint Mary's College of California 12

## 22 yes Saint Mary's College of California 2

## 23 no SmallPrivate 17

## 24 yes SmallPrivate 1

## 25 no Tokyo Woman's Christian University, Japan 19

## 26 yes Tokyo Woman's Christian University, Japan 2

## 27 no Yeshiva University 2

## 28 yes Yeshiva University 4

# calculate the percentages

df1 <- ddply(df1, .(almaMater), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df1 <- ddply(df1, .(almaMater), transform, pos = (cumsum(count) - 0.5 \* count))

df1$label <- paste0(sprintf("%.0f", df1$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df1, aes(x = almaMater, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Alma Mater')

#People from the Ladies group and the small private schools are slightly more likely to match

# create a dataframe

df2 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$field))

names(df2) <- c('match', 'fieldOfStudy', 'count')

df2

## match fieldOfStudy count

## 1 no Law 529

## 2 yes Law 136

## 3 no Math 186

## 4 yes Math 21

## 5 no Social Science 568

## 6 yes Social Science 128

## 7 no Medical Science 97

## 8 yes Medical Science 46

## 9 no Engineering 747

## 10 yes Engineering 117

## 11 no English 277

## 12 yes English 48

## 13 no History/Philosophy 201

## 14 yes History/Philosophy 40

## 15 no Business 1587

## 16 yes Business 338

## 17 no Education 531

## 18 yes Education 95

## 19 no Science 832

## 20 yes Science 161

## 21 no Social Work 418

## 22 yes Social Work 50

## 23 no Undecided 18

## 24 yes Undecided 1

## 25 no PolySci 590

## 26 yes PolySci 119

## 27 no Film 103

## 28 yes Film 23

## 29 no Fine Arts 165

## 30 yes Fine Arts 22

## 31 no Languages 28

## 32 yes Languages 12

## 33 no Architecture 9

## 34 yes Architecture 1

## 35 no Other 44

## 36 yes Other 8

# calculate the percentages

df2 <- ddply(df2, .(fieldOfStudy), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df2 <- ddply(df2, .(fieldOfStudy), transform, pos = (cumsum(count) - 0.5 \* count))

df2$label <- paste0(sprintf("%.0f", df2$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df2, aes(x = fieldOfStudy, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level with Different Fields of Study')

#The different fields of study yield more diversity, poor math majors....also medical majors, lawyers, and people

#who can speak foreign languages are very popular.

# create a dataframe

df3 <- data.frame(table(TidywalmaMater$matchYN, TidywalmaMater$brains))

names(df3) <- c('match', 'intelligence', 'count')

df3

## match intelligence count

## 1 0 4351

## 2 1 894

## 3 0 1,011.00 17

## 4 1 1,011.00 2

## 5 0 1,014.00 15

## 6 1 1,014.00 3

## 7 0 1,030.00 32

## 8 1 1,030.00 8

## 9 0 1,034.00 14

## 10 1 1,034.00 0

## 11 0 1,050.00 21

## 12 1 1,050.00 0

## 13 0 1,060.00 14

## 14 1 1,060.00 4

## 15 0 1,070.00 9

## 16 1 1,070.00 0

## 17 0 1,080.00 5

## 18 1 1,080.00 1

## 19 0 1,090.00 27

## 20 1 1,090.00 2

## 21 0 1,092.00 11

## 22 1 1,092.00 3

## 23 0 1,097.00 14

## 24 1 1,097.00 0

## 25 0 1,099.00 9

## 26 1 1,099.00 0

## 27 0 1,100.00 21

## 28 1 1,100.00 0

## 29 0 1,105.00 12

## 30 1 1,105.00 7

## 31 0 1,110.00 23

## 32 1 1,110.00 7

## 33 0 1,130.00 39

## 34 1 1,130.00 4

## 35 0 1,134.00 6

## 36 1 1,134.00 1

## 37 0 1,140.00 17

## 38 1 1,140.00 1

## 39 0 1,149.00 19

## 40 1 1,149.00 2

## 41 0 1,155.00 22

## 42 1 1,155.00 8

## 43 0 1,157.00 17

## 44 1 1,157.00 1

## 45 0 1,159.00 19

## 46 1 1,159.00 1

## 47 0 1,160.00 37

## 48 1 1,160.00 5

## 49 0 1,178.00 7

## 50 1 1,178.00 3

## 51 0 1,180.00 10

## 52 1 1,180.00 8

## 53 0 1,185.00 12

## 54 1 1,185.00 2

## 55 0 1,188.00 67

## 56 1 1,188.00 13

## 57 0 1,191.00 7

## 58 1 1,191.00 1

## 59 0 1,200.00 28

## 60 1 1,200.00 6

## 61 0 1,206.00 30

## 62 1 1,206.00 3

## 63 0 1,210.00 45

## 64 1 1,210.00 16

## 65 0 1,212.00 14

## 66 1 1,212.00 0

## 67 0 1,214.00 8

## 68 1 1,214.00 1

## 69 0 1,215.00 8

## 70 1 1,215.00 1

## 71 0 1,220.00 21

## 72 1 1,220.00 0

## 73 0 1,227.00 22

## 74 1 1,227.00 0

## 75 0 1,230.00 6

## 76 1 1,230.00 0

## 77 0 1,239.00 22

## 78 1 1,239.00 0

## 79 0 1,242.00 33

## 80 1 1,242.00 3

## 81 0 1,250.00 48

## 82 1 1,250.00 5

## 83 0 1,258.00 31

## 84 1 1,258.00 11

## 85 0 1,260.00 104

## 86 1 1,260.00 29

## 87 0 1,267.00 5

## 88 1 1,267.00 1

## 89 0 1,270.00 10

## 90 1 1,270.00 0

## 91 0 1,280.00 21

## 92 1 1,280.00 1

## 93 0 1,290.00 165

## 94 1 1,290.00 25

## 95 0 1,308.00 16

## 96 1 1,308.00 5

## 97 0 1,309.00 115

## 98 1 1,309.00 23

## 99 0 1,310.00 69

## 100 1 1,310.00 11

## 101 0 1,320.00 40

## 102 1 1,320.00 2

## 103 0 1,330.00 72

## 104 1 1,330.00 12

## 105 0 1,331.00 17

## 106 1 1,331.00 1

## 107 0 1,340.00 119

## 108 1 1,340.00 27

## 109 0 1,341.00 15

## 110 1 1,341.00 3

## 111 0 1,360.00 97

## 112 1 1,360.00 19

## 113 0 1,365.00 18

## 114 1 1,365.00 1

## 115 0 1,370.00 12

## 116 1 1,370.00 7

## 117 0 1,380.00 50

## 118 1 1,380.00 11

## 119 0 1,400.00 341

## 120 1 1,400.00 62

## 121 0 1,402.00 17

## 122 1 1,402.00 7

## 123 0 1,410.00 25

## 124 1 1,410.00 2

## 125 0 1,430.00 215

## 126 1 1,430.00 47

## 127 0 1,450.00 134

## 128 1 1,450.00 29

## 129 0 1,460.00 79

## 130 1 1,460.00 21

## 131 0 1,470.00 10

## 132 1 1,470.00 0

## 133 0 1,490.00 19

## 134 1 1,490.00 2

## 135 0 914.00 6

## 136 1 914.00 2

## 137 0 990.00 17

## 138 1 990.00 3

# calculate the percentages

df3 <- ddply(df3, .(intelligence), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df3 <- ddply(df3, .(intelligence), transform, pos = (cumsum(count) - 0.5 \* count))

df3$label <- paste0(sprintf("%.0f", df3$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df3, aes(x = intelligence, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Intelligence')

#not too much going on here, at first glance, high portion of na. probably should get rid of this one.

#I just came back to look at it with a different grouping. this time I used 1,000s, 1100s, etc. going to look at new graph...

#summary(TidywalmaMater$brains)

#levels(TidywalmaMater$brains)

#levels(TidywalmaMater$brains)[1] <- "na"

#levels(TidywalmaMater$brains)[2:13] <- "1,000s"

#levels(TidywalmaMater$brains)[3:18] <- "1,100s"

#levels(TidywalmaMater$brains)[4:21] <- "1,200s"

#levels(TidywalmaMater$brains)[5:16] <- "1,300s"

#levels(TidywalmaMater$brains)[6:13] <- "1,400s"

#levels(TidywalmaMater$brains)[7:8] <- "1,000s"

#It's a little better....still not very good. Still ok wih discarding this. I feel this should be more important, but I guess

#it isn't...Too bad there is no standard rating for looks available.......

# create a dataframe

df4 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$goal))

names(df4) <- c('match', 'Goal', 'count')

df4

## match Goal count

## 1 no Seemed like a fun night out 2843

## 2 yes Seemed like a fun night out 583

## 3 no To meet new people 2528

## 4 yes To meet new people 484

## 5 no To get a date 531

## 6 yes To get a date 100

## 7 no Looking for a serious relationship 250

## 8 yes Looking for a serious relationship 51

## 9 no To say I did it 425

## 10 yes To say I did it 85

## 11 no Other 357

## 12 yes Other 62

# calculate the percentages

df4 <- ddply(df4, .(Goal), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df4 <- ddply(df4, .(Goal), transform, pos = (cumsum(count) - 0.5 \* count))

df4$label <- paste0(sprintf("%.0f", df4$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df4, aes(x = Goal, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Goal')

#not too much going on here, probably noise/one to get rid of

# create a dataframe

df5 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$dateFreq))

names(df5) <- c('match', 'dateFrequency', 'count')

df5

## match dateFrequency count

## 1 no Several times a week 65

## 2 yes Several times a week 29

## 3 no Twice a week 243

## 4 yes Twice a week 65

## 5 no Once a week 624

## 6 yes Once a week 159

## 7 no Twice a month 1675

## 8 yes Twice a month 365

## 9 no Once a month 1291

## 10 yes Once a month 237

## 11 no Several times a year 1797

## 12 yes Several times a year 297

## 13 no Almost Never 1222

## 14 yes Almost Never 212

# calculate the percentages

df5 <- ddply(df5, .(dateFrequency), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df5 <- ddply(df5, .(dateFrequency), transform, pos = (cumsum(count) - 0.5 \* count))

df5$label <- paste0(sprintf("%.0f", df5$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df5, aes(x = dateFrequency, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Date Frequency')

#this one might be good, nice variance here compared to some of the others

# create a dataframe

df6 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$socialScore))

names(df6) <- c('match', 'Sociability', 'count')

df6

## match Sociability count

## 1 no Several times a week 2103

## 2 yes Several times a week 507

## 3 no Twice a week 2511

## 4 yes Twice a week 479

## 5 no Once a week 1660

## 6 yes Once a week 289

## 7 no Twice a month 393

## 8 yes Twice a month 57

## 9 no Once a month 145

## 10 yes Once a month 19

## 11 no Several times a year 86

## 12 yes Several times a year 13

## 13 no Almost never 36

## 14 yes Almost never 1

# calculate the percentages

df6 <- ddply(df6, .(Sociability), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df6 <- ddply(df6, .(Sociability), transform, pos = (cumsum(count) - 0.5 \* count))

df6$label <- paste0(sprintf("%.0f", df6$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df6, aes(x = Sociability, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Social Score')

#This one has some variance but I worry that social score and date frequency might be correlated, so this one

#which is not as good as the other one, may need to go.

# create a dataframe

df7 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$age))

names(df7) <- c('match', 'Age', 'count')

df7

## match Age count

## 1 no 18 10

## 2 yes 18 0

## 3 no 19 15

## 4 yes 19 5

## 5 no 20 41

## 6 yes 20 14

## 7 no 21 226

## 8 yes 21 65

## 9 no 22 561

## 10 yes 22 94

## 11 no 23 750

## 12 yes 23 144

## 13 no 24 740

## 14 yes 24 123

## 15 no 25 641

## 16 yes 25 196

## 17 no 26 723

## 18 yes 26 146

## 19 no 27 891

## 20 yes 27 168

## 21 no 28 616

## 22 yes 28 130

## 23 no 29 496

## 24 yes 29 93

## 25 no 30 491

## 26 yes 30 83

## 27 no 31 107

## 28 yes 31 18

## 29 no 32 189

## 30 yes 32 21

## 31 no 33 141

## 32 yes 33 20

## 33 no 34 132

## 34 yes 34 20

## 35 no 35 50

## 36 yes 35 10

## 37 no 36 41

## 38 yes 36 4

## 39 no 37 5

## 40 yes 37 0

## 41 no 38 17

## 42 yes 38 2

## 43 no 39 16

## 44 yes 39 2

## 45 no 42 16

## 46 yes 42 4

## 47 no 55 6

## 48 yes 55 0

# calculate the percentages

df7 <- ddply(df7, .(Age), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df7 <- ddply(df7, .(Age), transform, pos = (cumsum(count) - 0.5 \* count))

df7$label <- paste0(sprintf("%.0f", df7$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df7, aes(x = Age, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Age of Respondent')

#This is probably the best one so far. Not too surprising really

# create a dataframe

df71 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$agePart))

names(df71) <- c('match', 'AgePart', 'count')

df71

## match AgePart count

## 1 no 18 9

## 2 yes 18 0

## 3 no 19 14

## 4 yes 19 5

## 5 no 20 40

## 6 yes 20 14

## 7 no 21 224

## 8 yes 21 65

## 9 no 22 557

## 10 yes 22 94

## 11 no 23 750

## 12 yes 23 144

## 13 no 24 740

## 14 yes 24 123

## 15 no 25 641

## 16 yes 25 196

## 17 no 26 723

## 18 yes 26 146

## 19 no 27 891

## 20 yes 27 168

## 21 no 28 616

## 22 yes 28 130

## 23 no 29 496

## 24 yes 29 93

## 25 no 30 491

## 26 yes 30 83

## 27 no 31 107

## 28 yes 31 18

## 29 no 32 189

## 30 yes 32 21

## 31 no 33 141

## 32 yes 33 20

## 33 no 34 132

## 34 yes 34 20

## 35 no 35 50

## 36 yes 35 10

## 37 no 36 41

## 38 yes 36 4

## 39 no 37 5

## 40 yes 37 0

## 41 no 38 17

## 42 yes 38 2

## 43 no 39 16

## 44 yes 39 2

## 45 no 42 16

## 46 yes 42 4

## 47 no 55 6

## 48 yes 55 0

# calculate the percentages

df71 <- ddply(df71, .(AgePart), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df71 <- ddply(df71, .(AgePart), transform, pos = (cumsum(count) - 0.5 \* count))

df71$label <- paste0(sprintf("%.0f", df71$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df71, aes(x = AgePart, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Age of Partner')

#if you aren't under 21 hope you are 25

# create a dataframe

df8 <- data.frame(table(TidydataPrePurge$matchYN, TidydataPrePurge$attitude))

names(df8) <- c('match', 'Attitude', 'count')

df8

## match Attitude count

## 1 no 1 101

## 2 yes 1 15

## 3 no 2 252

## 4 yes 2 45

## 5 no 3 604

## 6 yes 3 102

## 7 no 4 673

## 8 yes 4 120

## 9 no 5 1701

## 10 yes 5 332

## 11 no 6 1653

## 12 yes 6 350

## 13 no 7 1230

## 14 yes 7 245

## 15 no 8 423

## 16 yes 8 90

## 17 no 9 177

## 18 yes 9 39

## 19 no 10 99

## 20 yes 10 26

# calculate the percentages

df8 <- ddply(df8, .(Attitude), transform, percent = count/sum(count) \* 100)

# format the labels and calculate their positions

df8 <- ddply(df8, .(Attitude), transform, pos = (cumsum(count) - 0.5 \* count))

df8$label <- paste0(sprintf("%.0f", df8$percent), "%")

# bar plot of counts by occupation with in group proportions

ggplot(df8, aes(x = Attitude, y = count, fill = match)) +

geom\_bar(stat = "identity") +

geom\_text(aes(y = pos, label = label), size = 2) +

ggtitle('Match Level by Expected Happiness')

# Not alot going on here good candidate for noise reduction....

#Home has too many factors I think, I'm going to get rid of it

dating$home <- NULL

dating$goal <- NULL

dating$socialScore <- NULL

dating$attitude<- NULL

dating$brains<-NULL

dating$swagLevel <- NULL

#After doing evaluation and backwards selection I am removing gender

dating$gender <- NULL

summary(dating)

## matchYN intRate agePart age

## no :6998 Min. :-0.830 Min. :18.00 Min. :18.00

## yes:1380 1st Qu.:-0.020 1st Qu.:24.00 1st Qu.:24.00

## Median : 0.210 Median :26.00 Median :26.00

## Mean : 0.196 Mean :26.36 Mean :26.36

## 3rd Qu.: 0.430 3rd Qu.:28.00 3rd Qu.:28.00

## Max. : 0.910 Max. :55.00 Max. :55.00

## NA's :158 NA's :104 NA's :95

## field almaMater dateFreq

## Business :1925 Other :4347 Several times a year:2094

## Science : 993 CaliSchool : 911 Twice a month :2040

## Engineering : 864 Ivy : 882 Once a month :1528

## PolySci : 709 European School: 835 Almost Never :1434

## Social Science: 696 BigPrivate : 544 Once a week : 783

## (Other) :3109 Asian School : 472 (Other) : 402

## NA's : 82 (Other) : 387 NA's : 97

#Spliting the data to test some models! Setting seed to ensure replicable results.

set.seed(22)

#partition

dset <-createDataPartition(dating$matchYN, p=4/5, list=F)

#double-check the partition size

nrow(dset) / nrow(dating)

## [1] 0.8000716

#create training/test data

trainDate <-dating[dset,]

testDate <-dating[-dset,]

# Function that returns Root Mean Squared Error

rmse <- function(error)

{

sqrt(mean(error^2))

}

# Function that returns Mean Absolute Error

mae <- function(error)

{

mean(abs(error))

}

logregmod <- glm(matchYN ~ ., data = trainDate, family = binomial('logit'))

# evaluation

m\_full <- logregmod # full model is the model just fitted

m\_null <- glm(matchYN ~ 1, data = trainDate, family = binomial('logit'))

# backward selection

step(m\_full, trace = F, scope = list(lower=formula(m\_null), upper=formula(m\_full)),

direction = 'backward')

##

## Call: glm(formula = matchYN ~ intRate + agePart + age + field + dateFreq,

## family = binomial("logit"), data = trainDate)

##

## Coefficients:

## (Intercept) intRate

## 0.528093 0.400570

## agePart age

## -0.024858 -0.032769

## fieldMath fieldSocial Science

## -0.825710 -0.001423

## fieldMedical Science fieldEngineering

## 0.800419 -0.508585

## fieldEnglish fieldHistory/Philosophy

## -0.239525 -0.173719

## fieldBusiness fieldEducation

## -0.118416 -0.215426

## fieldScience fieldSocial Work

## -0.190384 -0.767086

## fieldUndecided fieldPolySci

## -1.326814 -0.254852

## fieldFilm fieldFine Arts

## -0.033671 -0.519253

## fieldLanguages fieldArchitecture

## 0.304198 -12.170416

## fieldOther dateFreqTwice a week

## -0.420903 -0.239209

## dateFreqOnce a week dateFreqTwice a month

## -0.348577 -0.410403

## dateFreqOnce a month dateFreqSeveral times a year

## -0.570246 -0.695127

## dateFreqAlmost Never

## -0.585520

##

## Degrees of Freedom: 6488 Total (i.e. Null); 6462 Residual

## (214 observations deleted due to missingness)

## Null Deviance: 5813

## Residual Deviance: 5703 AIC: 5757

step(logregmod, direction="backward")

## Start: AIC=5764.5

## matchYN ~ intRate + agePart + age + field + almaMater + dateFreq

##

## Df Deviance AIC

## - almaMater 13 5702.8 5756.8

## <none> 5684.5 5764.5

## - agePart 1 5690.0 5768.0

## - dateFreq 6 5701.9 5769.9

## - age 1 5693.1 5771.1

## - intRate 1 5695.0 5773.0

## - field 17 5747.9 5793.9

##

## Step: AIC=5756.81

## matchYN ~ intRate + agePart + age + field + dateFreq

##

## Df Deviance AIC

## <none> 5702.8 5756.8

## - agePart 1 5709.2 5761.2

## - dateFreq 6 5721.2 5763.2

## - age 1 5712.6 5764.6

## - intRate 1 5715.1 5767.1

## - field 17 5766.9 5786.9

##

## Call: glm(formula = matchYN ~ intRate + agePart + age + field + dateFreq,

## family = binomial("logit"), data = trainDate)

##

## Coefficients:

## (Intercept) intRate

## 0.528093 0.400570

## agePart age

## -0.024858 -0.032769

## fieldMath fieldSocial Science

## -0.825710 -0.001423

## fieldMedical Science fieldEngineering

## 0.800419 -0.508585

## fieldEnglish fieldHistory/Philosophy

## -0.239525 -0.173719

## fieldBusiness fieldEducation

## -0.118416 -0.215426

## fieldScience fieldSocial Work

## -0.190384 -0.767086

## fieldUndecided fieldPolySci

## -1.326814 -0.254852

## fieldFilm fieldFine Arts

## -0.033671 -0.519253

## fieldLanguages fieldArchitecture

## 0.304198 -12.170416

## fieldOther dateFreqTwice a week

## -0.420903 -0.239209

## dateFreqOnce a week dateFreqTwice a month

## -0.348577 -0.410403

## dateFreqOnce a month dateFreqSeveral times a year

## -0.570246 -0.695127

## dateFreqAlmost Never

## -0.585520

##

## Degrees of Freedom: 6488 Total (i.e. Null); 6462 Residual

## (214 observations deleted due to missingness)

## Null Deviance: 5813

## Residual Deviance: 5703 AIC: 5757

rmse(logregmod$residuals)

## [1] 2.81548

mae(logregmod$residuals)

## [1] 2.000372

1- with(logregmod, deviance/null.deviance)

## [1] 0.02218398

# RMSE=2.816 MAE=2 R^2=.023

logregmod.pred.test <- predict(logregmod, testDate, type = "response", positive=2)

# If logregmod.pred.test exceeds threshold of 0.5, yes else no

y\_or\_n <- ifelse(logregmod.pred.test>0.5, "yes", "no")

# Convert to factor: p\_class

p\_class <- factor( y\_or\_n, levels = levels(testDate$matchYN))

# Create confusion matrix

confusionMatrix(p\_class, testDate$matchYN)

## Confusion Matrix and Statistics

##

## Reference

## Prediction no yes

## no 1352 269

## yes 1 1

##

## Accuracy : 0.8336

## 95% CI : (0.8146, 0.8515)

## No Information Rate : 0.8336

## P-Value [Acc > NIR] : 0.5162

##

## Kappa : 0.0049

##

## Mcnemar's Test P-Value : <2e-16

##

## Sensitivity : 0.999261

## Specificity : 0.003704

## Pos Pred Value : 0.834053

## Neg Pred Value : 0.500000

## Prevalence : 0.833641

## Detection Rate : 0.833025

## Detection Prevalence : 0.998768

## Balanced Accuracy : 0.501482

##

## 'Positive' Class : no

##

#This model is accurate but it does not do well guessing the yes responses. It only guessed 2 out of 268.

summary(dating$matchYN)

## no yes

## 6998 1380

#Fun fact being Pre Med is the best way to get people to match with you

#the worst thing you could do is study Architecture, although that was a small number of people.

#the worst thing a significant number of people do is be undecided on thier major.

#These are the features we are going to work with

#This data set is dominated by "no" matches, I'm going to try to fix that by over/under sampling.

#The problem is the model is seeking to be accurate, and the best way to be accurate is to simply guess no for the

#match response. So what I'm going to do is provide more samples of yes to train the model to pick all the yes answers.

#Since I am trying to find the yes answers as my goal this method should help. I am making sure to only oversample the training

#set so information from the test set doesn't make its way into the sample. I would like to look into SMOTE but we haven't covered that.

library(ROSE)

## Warning: package 'ROSE' was built under R version 3.5.3

## Loaded ROSE 0.0-3

over <- ovun.sample(matchYN ~., data=trainDate, method="over")$data

under <- ovun.sample(matchYN~. , data=trainDate, method = "under")$data

both <- ovun.sample(matchYN~. , data=trainDate, method = "both")$data

logregmodover <- glm(matchYN ~ ., data = over, family = binomial('logit'))

rmse(logregmodover$residuals)

## [1] 2.048859

mae(logregmodover$residuals)

## [1] 2.000997

1- with(logregmodover, deviance/null.deviance)

## [1] 0.02532499

# RMSE= 2.04 MAE= 2 R^2=.03

logregmodover.pred.test.train <- predict(logregmodover, trainDate, type = "response", positive=2)

# If logregmod.pred.test exceeds threshold of 0.5, yes else no

y\_or\_nover.train <- ifelse(logregmodover.pred.test.train>0.5, "yes", "no")

# Convert to factor: p\_class

p\_classover.train <- factor( y\_or\_nover.train, levels = levels(trainDate$matchYN))

# Create confusion matrix

confusionMatrix(p\_classover.train, trainDate$matchYN)

## Confusion Matrix and Statistics

##

## Reference

## Prediction no yes

## no 3050 453

## yes 2368 618

##

## Accuracy : 0.5653

## 95% CI : (0.5531, 0.5774)

## No Information Rate : 0.835

## P-Value [Acc > NIR] : 1

##

## Kappa : 0.0815

##

## Mcnemar's Test P-Value : <2e-16

##

## Sensitivity : 0.5629

## Specificity : 0.5770

## Pos Pred Value : 0.8707

## Neg Pred Value : 0.2070

## Prevalence : 0.8350

## Detection Rate : 0.4700

## Detection Prevalence : 0.5398

## Balanced Accuracy : 0.5700

##

## 'Positive' Class : no

##

632/(436+632)

## [1] 0.5917603

logregmodover.pred.test <- predict(logregmodover, testDate, type = "response", positive=2)

# If logregmod.pred.test exceeds threshold of 0.5, yes else no

y\_or\_nover <- ifelse(logregmodover.pred.test>0.5, "yes", "no")

# Convert to factor: p\_class

p\_classover <- factor( y\_or\_nover, levels = levels(testDate$matchYN))

# Create confusion matrix

confusionMatrix(p\_classover, testDate$matchYN)

## Confusion Matrix and Statistics

##

## Reference

## Prediction no yes

## no 768 122

## yes 585 148

##

## Accuracy : 0.5644

## 95% CI : (0.5399, 0.5887)

## No Information Rate : 0.8336

## P-Value [Acc > NIR] : 1

##

## Kappa : 0.0687

##

## Mcnemar's Test P-Value : <2e-16

##

## Sensitivity : 0.5676

## Specificity : 0.5481

## Pos Pred Value : 0.8629

## Neg Pred Value : 0.2019

## Prevalence : 0.8336

## Detection Rate : 0.4732

## Detection Prevalence : 0.5484

## Balanced Accuracy : 0.5579

##

## 'Positive' Class : no

##

145/(145+125)

## [1] 0.537037

#got 54% of the yeses

#The accuracy has gone down but the model had correctly predicted over half of the yeses

#instead of virtually none of them.

#logregmodunder <- glm(matchYN ~ ., data = under, family = binomial('logit'))

#rmse(logregmodunder$residuals)

#mae(logregmodunder$residuals)

#1- with(logregmodunder, deviance/null.deviance)

# RMSE= 2.05 MAE= 2 R^2=.03

#logregmodunder.pred.test <- predict(logregmodunder, testDate, type = "response", positive=2)

#y\_or\_nunder <- ifelse(logregmodunder.pred.test>0.5, "yes", "no")

# Convert to factor: p\_class

#p\_classunder <- factor( y\_or\_nunder, levels = levels(testDate$matchYN))

# Create confusion matrix

#confusionMatrix(p\_classunder, testDate$matchYN)

logregmodboth <- glm(matchYN ~ ., data = both, family = binomial('logit'))

rmse(logregmodboth$residuals)

## [1] 2.036417

mae(logregmodboth$residuals)

## [1] 1.996579

1- with(logregmodboth, deviance/null.deviance)

## [1] 0.02826916

# RMSE= 2.05 MAE= 2 R^2=.03

logregmodboth.pred.test <- predict(logregmodboth, testDate, type = "response", positive=2)

y\_or\_nboth <- ifelse(logregmodboth.pred.test>0.5, "yes", "no")

# Convert to factor: p\_class

p\_classboth <- factor( y\_or\_nboth, levels = levels(testDate$matchYN))

# Create confusion matrix

confusionMatrix(p\_classboth, testDate$matchYN)

## Confusion Matrix and Statistics

##

## Reference

## Prediction no yes

## no 796 129

## yes 557 141

##

## Accuracy : 0.5773

## 95% CI : (0.5529, 0.6015)

## No Information Rate : 0.8336

## P-Value [Acc > NIR] : 1

##

## Kappa : 0.0676

##

## Mcnemar's Test P-Value : <2e-16

##

## Sensitivity : 0.5883

## Specificity : 0.5222

## Pos Pred Value : 0.8605

## Neg Pred Value : 0.2020

## Prevalence : 0.8336

## Detection Rate : 0.4904

## Detection Prevalence : 0.5699

## Balanced Accuracy : 0.5553

##

## 'Positive' Class : no

##

142/(142+128)

## [1] 0.5259259

#got 53% of the yeses

library(rpart)

tree2 <- rpart(matchYN ~ ., data = trainDate, method = 'class', cp = 1e-3)

tree2.pred.prob <- predict(tree2, newdata = testDate, type = 'prob')

tree2.pred <- predict(tree2, newdata = testDate, type = 'class')

# confusion matrix

tb2 <- table(tree2.pred, testDate$matchYN)

tb2

##

## tree2.pred no yes

## no 1393 267

## yes 6 9

#the regular cart model did better than the regular GLM time to try it with the different samples

tree2over <- rpart(matchYN ~ ., data = over, method = 'class', cp = 1e-3)

tree2over.pred.prob <- predict(tree2over, newdata = testDate, type = 'prob')

tree2over.pred <- predict(tree2over, newdata = testDate, type = 'class')

# confusion matrix

tb2 <- table(tree2over.pred, testDate$matchYN)

tb2

##

## tree2over.pred no yes

## no 928 155

## yes 471 121

134/(134+142)

## [1] 0.4855072

#this got 49% of the yeses

tree2both <- rpart(matchYN ~ ., data = both, method = 'class', cp = 1e-3)

tree2both.pred.prob <- predict(tree2both, newdata = testDate, type = 'prob')

tree2both.pred <- predict(tree2both, newdata = testDate, type = 'class')

# confusion matrix

tb2 <- table(tree2both.pred, testDate$matchYN)

tb2

##

## tree2both.pred no yes

## no 850 143

## yes 549 133

113/(113+163)

## [1] 0.4094203

#this got 41% of the yeses

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.5.2

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##

## Attaching package: 'randomForest'

## The following object is masked from 'package:ranger':

##

## importance

## The following object is masked from 'package:ggplot2':

##

## margin

ControlParameters <-trainControl(method="cv",

number=5,

savePredictions = TRUE,

classProbs = TRUE)

parameterGrid <-expand.grid(mtry=c(2,3,4))

modelRandomover <-train(matchYN~.,

data=over,

method="rf",

trControl=ControlParameters,

tuneGrid=parameterGrid)

rfover.pred.train <- predict(modelRandomover, trainDate, type = "prob", positive=2)

y\_or\_nrfover.train <- ifelse(rfover.pred.train>0.5, "yes", "no")

# Convert to factor: p\_class

p\_classrfover.train <- factor( y\_or\_nrfover.train, levels = levels(trainDate$matchYN))

# Create confusion matrix

confusionMatrix(p\_classover.train, trainDate$matchYN, positive = "yes")

## Confusion Matrix and Statistics

##

## Reference

## Prediction no yes

## no 3050 453

## yes 2368 618

##

## Accuracy : 0.5653

## 95% CI : (0.5531, 0.5774)

## No Information Rate : 0.835

## P-Value [Acc > NIR] : 1

##

## Kappa : 0.0815

##

## Mcnemar's Test P-Value : <2e-16

##

## Sensitivity : 0.57703

## Specificity : 0.56294

## Pos Pred Value : 0.20697

## Neg Pred Value : 0.87068

## Prevalence : 0.16505

## Detection Rate : 0.09524

## Detection Prevalence : 0.46016

## Balanced Accuracy : 0.56998

##

## 'Positive' Class : yes

##

632/(439+632)

## [1] 0.5901027

rfover.pred.test <- predict(modelRandomover, testDate, type = "prob", positive=2)

y\_or\_nrfover <- ifelse(logregmodover.pred.test>0.5, "yes", "no")

# Convert to factor: p\_class

p\_classrfover <- factor( y\_or\_nrfover, levels = levels(testDate$matchYN))

# Create confusion matrix

confusionMatrix(p\_classover, testDate$matchYN, positive = "yes")

## Confusion Matrix and Statistics

##

## Reference

## Prediction no yes

## no 768 122

## yes 585 148

##

## Accuracy : 0.5644

## 95% CI : (0.5399, 0.5887)

## No Information Rate : 0.8336

## P-Value [Acc > NIR] : 1

##

## Kappa : 0.0687

##

## Mcnemar's Test P-Value : <2e-16

##

## Sensitivity : 0.54815

## Specificity : 0.56763

## Pos Pred Value : 0.20191

## Neg Pred Value : 0.86292

## Prevalence : 0.16636

## Detection Rate : 0.09119

## Detection Prevalence : 0.45163

## Balanced Accuracy : 0.55789

##

## 'Positive' Class : yes

##

145/(145+125)

## [1] 0.537037

rf3 <- randomForest(matchYN ~ ., data = trainDate, ntree = 1000, na.action = na.omit)

rf3.pred.prob <- predict(rf3, newdata = testDate, type = 'prob')

rf3.pred <- predict(rf3, newdata = testDate, type = 'class')

# confusion matrix

tb3 <- table(rf3.pred, testDate$matchYN)

tb3

##

## rf3.pred no yes

## no 1323 251

## yes 30 19

#It got 18 out of 270 yeses

rfover <- randomForest(matchYN ~ ., data = over, ntree = 1000, na.action = na.omit)

rfover.pred.prob <- predict(rfover, newdata = testDate, type = 'prob')

rfover.pred <- predict(rfover, newdata = testDate, type = 'class')

# confusion matrix

tb3over <- table(rfover.pred, testDate$matchYN)

tb3over

##

## rfover.pred no yes

## no 1169 209

## yes 184 61

#it got 66/270

rfboth <- randomForest(matchYN ~ ., data = both, ntree = 1000, na.action = na.omit)

rfboth.pred.prob <- predict(rfboth, newdata = testDate, type = 'prob')

rfboth.pred <- predict(rfboth, newdata = testDate, type = 'class')

# confusion matrix

tb3both <- table(rfboth.pred, testDate$matchYN)

tb3both

##

## rfboth.pred no yes

## no 1048 162

## yes 305 108

#lrmover <-train(matchYN~.,

# data=over,

# method="glm",

# trControl=ControlParameters)

#lrmover.pred.test <- predict(lrmover, testDate, type = "prob", positive=2)

### If logregmod.pred.test exceeds threshold of 0.5, yes else no

#y\_or\_nlrmover <- ifelse(lrmover.pred.test$yes>0.5, "yes", "no")

### Convert to factor: p\_class

#p\_classlrmover <- factor( y\_or\_nlrmover, levels = levels(testDate$matchYN))

### Create confusion matrix

#confusionMatrix(p\_classlrmover, testDate$matchYN)