W241 Final Project EDA

Kalvin Kao

December 16, 2017

```r
# load packages
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.3.3
## Loading required package: zoo
## Warning: package 'zoo' was built under R version 3.3.3
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 3.3.3
```

```r
library(multiwayvcov)
```

```
## Warning: package 'multiwayvcov' was built under R version 3.3.3
```

```r
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.3.3
```

```r
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 3.3.3
```

```r
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.3.3
```

```r
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```r
robustSEs <- function(my.model){
  my.model$vcovHC <- vcovHC(my.model)
  my.model.summary <- coeftest(my.model, my.model$vcovHC)
  return(my.model.summary)
}
```

```r
#setwd('C:/MIDS/W241/final_project/Analysis')
df <- fread("Dating_experiment-Final_Project_DataFV.csv")
head(df)
```

```
##                                                                                                                          pro
## 1: https://images-ssl.gotinder.com/5a11e19e8802dc4401476da7/1080x1080_09ae0c2d-9adc-405a-866d-86a4ef
## 2: https://images-ssl.gotinder.com/59f5da9e1d0f17950c2ddba0/1080x1080_c3716765-d582-4486-b968-6fe99b
## 3:           https://images-ssl.gotinder.com/558e1372af6b25c50e064d2c/6642e34d-a032-483b-abe4-58a5ab
## 4:           https://images-ssl.gotinder.com/58ecbfb9fb6535500bc827f3/0fea504e-041f-452b-9bd4-475b25
## 5: https://images-ssl.gotinder.com/5a05ee4ab56f6be5038b20d0/1080x1080_9a296440-7221-438d-8b13-d90bff
## 6: https://images-ssl.gotinder.com/5a0e90a5612b424b01607de6/1080x1080_778dae6d-d52e-4518-981d-8c96b5
##                                   name age                       detail_1
## 1: \\uc0\\u1052 \\u1080 \\u1096 \\u1072  31
## 2:                                   A  26
## 3:                              Aakash  25
## 4:                              Aamash  25          103 Instagram Photos
## 5:                               Aaron  26 Community Health Worker (CHW)
## 6:                               Aaron  27
##                              detail_2 Matched swipes num_details
## 1:                                          0      1           0
## 2:                                          0      1           0
## 3:                                          0      1           0
## 4:                                          0      1           1
## 5: UIC Jane Addams College of Social Work       0      1           2
## 6:                                          1      1           0
##    num_ig school job unknown noedu bs md phd associates female male
## 1:     NA      0   0       0     0  1  0   0          0      1    0
## 2:     NA      0   0       0     0  1  0   0          0      1    0
## 3:     NA      0   0       0     0  1  0   0          0      1    0
## 4:    103      0   0       0     0  1  0   0          0      1    0
## 5:     NA      1   0       1     0  1  0   0          0      1    0
## 6:     NA      0   0       0     0  1  0   0          0      1    0
##    chicago houston losangeles newyork philadelphia phoenix sanantonio
## 1:       0       0          1       0            0       0          0
## 2:       0       0          0       0            0       0          0
## 3:       1       0          0       0            0       0          0
## 4:       0       0          0       1            0       0          0
## 5:       1       0          0       0            0       0          0
## 6:       0       0          0       0            0       1          0
##    sandiego
## 1:        0
## 2:        1
## 3:        0
## 4:        0
## 5:        0
## 6:        0
```

```
summary(df)
```

```
##  profile_image          name                 age          detail_1
##  Length:6270        Length:6270        Min.   :24.00   Length:6270
##  Class :character   Class :character   1st Qu.:25.00   Class :character
##  Mode  :character   Mode  :character   Median :27.00   Mode  :character
##                                        Mean   :27.46
##                                        3rd Qu.:29.00
##                                        Max.   :33.00
##                                        NA's   :79
##    detail_2           Matched            swipes         num_details
##  Length:6270        Min.   :0.00000   Min.   : 1.000   Min.   :0.000
```

```
##  Class :character   1st Qu.:0.00000   1st Qu.: 1.000   1st Qu.:1.000
##  Mode  :character   Median :0.00000   Median : 1.000   Median :1.000
##                     Mean   :0.09316   Mean   : 1.034   Mean   :1.183
##                     3rd Qu.:0.00000   3rd Qu.: 1.000   3rd Qu.:2.000
##                     Max.   :1.00000   Max.   :52.000   Max.   :2.000
##                     NA's   :1
##      num_ig          school            job            unknown
##  Min.   :   2.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  93.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 236.0   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   : 422.6   Mean   :0.5018   Mean   :0.1834   Mean   :0.3437
##  3rd Qu.: 522.2   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :5160.0   Max.   :1.0000   Max.   :1.0000   Max.   :2.0000
##  NA's   :5406
##      noedu              bs               md              phd
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.2534   Mean   :0.2544   Mean   :0.2405   Mean   :0.2517
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##    associates     female            male           chicago
##  Min.   :0    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0    Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :0    Mean   :0.4938   Mean   :0.5062   Mean   :0.1252
##  3rd Qu.:0    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :0    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##     houston         losangeles        newyork        philadelphia
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1265   Mean   :0.1265   Mean   :0.1265   Mean   :0.1239
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##     phoenix         sanantonio       sandiego
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.1266   Mean   :0.118   Mean   :0.1268
##  3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##
```

EDA

The randomization of our treatment assignment was dependent on Tinder's selection of suitors. Since Tinder's algorithm for this selection is unknown and potentially complex, balance across the available covariates was examined to ensure that the experiment yields an apples-to-apples comparison.

Missing Instagram Information

Like school and job information, instagram information was missing from many of our suitor profiles. However, it was easily detected when present in the source code since the format for an instagram detail was consistent across suitor profiles (of the form "X Instagram Photos"). One technical detail of the structure of suitor source codes is that school and job information took priority– an instagram detail is only present when less than two school or job details are present. Therefore, whether or not a profile's source code contains an instagram detail could be used to represent the amount of information a suitor chose to include in his or her profile– the missingness of instagram details is thus checked for balance across experimental conditions. Table XXX below shows the count of profiles with instagram information detected between the male and female test profiles.

% latex table generated in R 3.3.2 by xtable 1.8-2 package % Sat Dec 16 22:05:10 2017

	Gender	Total Profiles	Profiles with Instagram	% Missing Instagram Info
1	Female	3096	439	85.8204134366925
2	Male	3174	425	86.6099558916194

Table XXX below shows the count of profiles with instagram information detected between all treatment conditions.

% latex table generated in R 3.3.2 by xtable 1.8-2 package % Sat Dec 16 22:05:10 2017

	Education Level	Total Profiles	Profiles with Instagram	% Missing Instagram Info
1	No Education	1589	262	83.5116425424795
2	BS	1595	185	88.4012539184953
3	MD	1508	213	85.8753315649867
4	PhD	1578	204	87.0722433460076

Table XXX below shows the count of profiles with instagram information detected between all testing locations.

% latex table generated in R 3.3.2 by xtable 1.8-2 package % Sat Dec 16 22:05:10 2017

	City	Total Profiles	Profiles with Instagram	% Missing Instagram Info
1	Chicago	785	90	88.5350318471338
2	Houston	793	91	88.5245901639344
3	Los Angeles	793	156	80.327868852459
4	New York	793	109	86.2547288776797
5	Phildelphia	777	108	86.1003861003861
6	Phoenix	794	106	86.6498740554156
7	San Antonio	740	80	89.1891891891892
8	San Diego	795	124	84.4025157232704

Table XXX below shows the results of the tests for a difference in average missingness of instagram information, by gender, treatment condition, and location. While some differences are statistically significant, we do not consider them practically significant, and the missingess of instagram information thus passes the balance check.

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 16, 2017 - 10:05:10 PM

Table 1: Comparison of IG Missingness

	Dependent variable:		
	Gender	Treatment	Location
	(1)	(2)	(3)
female	0.008 (0.009)		
bs		-0.049 (0.012)***	
md		-0.024 (0.013)*	
phd		-0.036 (0.013)***	
houston			0.0001 (0.016)
losangeles			0.082 (0.018)***
newyork			0.023 (0.017)
philadelphia			0.024 (0.017)
phoenix			0.019 (0.017)
sanantonio			-0.007 (0.016)
sandiego			0.041 (0.017)**
Constant	0.134 (0.006)***	0.165 (0.009)***	0.115 (0.011)***
Note:			*p$<$0.1; **p$<$0.05; ***p$<$0.01

Missing Age Values

Table XXX below shows the count of profiles that contain the suitor's age, between males and females.

% latex table generated in R 3.3.2 by xtable 1.8-2 package % Sat Dec 16 22:05:10 2017

	Gender	Total Profiles	Profiles with Age	% Missing Age
1	Female	3096	3035	1.97028423772609
2	Male	3174	3156	0.567107750472595

Table XXX below shows the count of profiles that contain the suitor's age, between all treatment conditions.

% latex table generated in R 3.3.2 by xtable 1.8-2 package % Sat Dec 16 22:05:10 2017

	Education Level	Total Profiles	Profiles with Age	% Missing Age
1	No Education	1589	1569	1.25865324103209
2	BS	1595	1571	1.50470219435737
3	MD	1508	1487	1.39257294429708
4	PhD	1578	1564	0.887198986058302

Table XXX below shows the count of profiles that contain the suitor's age, between all testing locations.

% latex table generated in R 3.3.2 by xtable 1.8-2 package % Sat Dec 16 22:05:10 2017

Table XXX below shows the results of the tests for a difference in average missingness of the suitor's age, by gender, treatment condition, and location. While some differences are statistically significant (between males and females, for example), we do not consider them practically significant, and the missingess of age information thus passes the balance check.

	City	Total Profiles	Profiles with Age	% Missing Age
1	Chicago	785	777	1.01910828025478
2	Houston	793	783	1.2610340479193
3	Los Angeles	793	775	2.26986128625473
4	New York	793	775	2.26986128625473
5	Phildelphia	777	772	0.643500643500639
6	Phoenix	794	791	0.377833753148615
7	San Antonio	740	735	0.67567567567568
8	San Diego	795	783	1.50943396226415

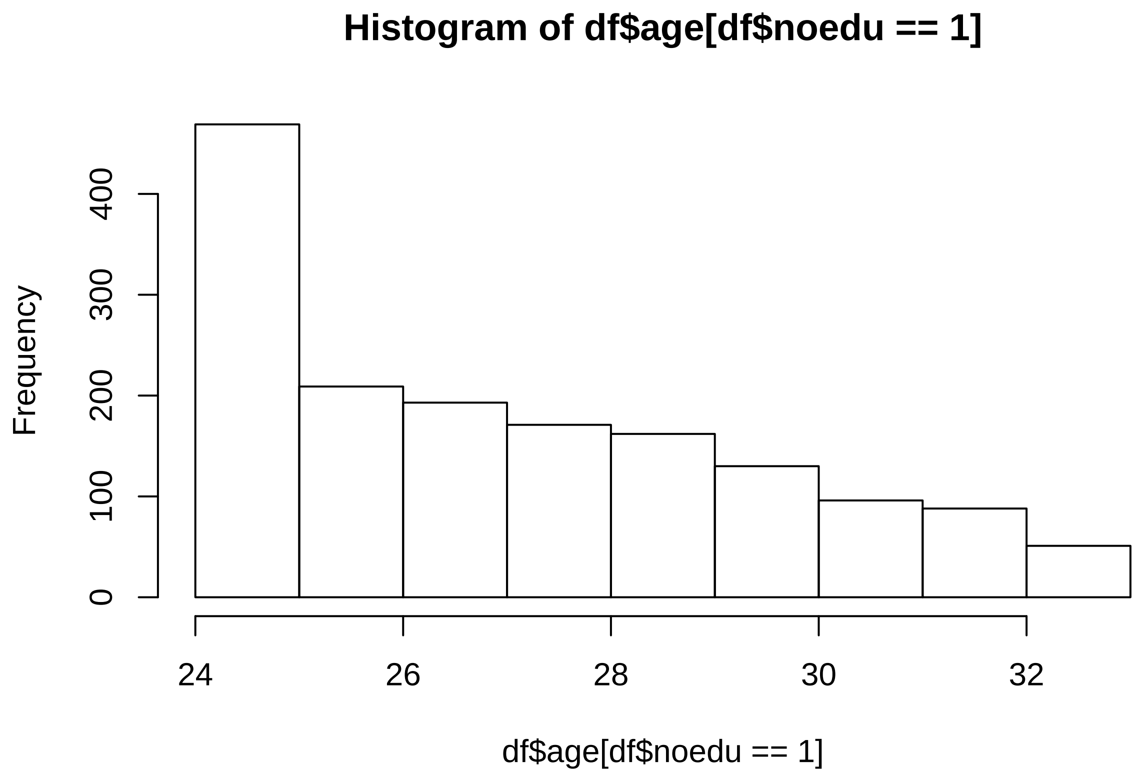
% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 16, 2017 - 10:05:11 PM

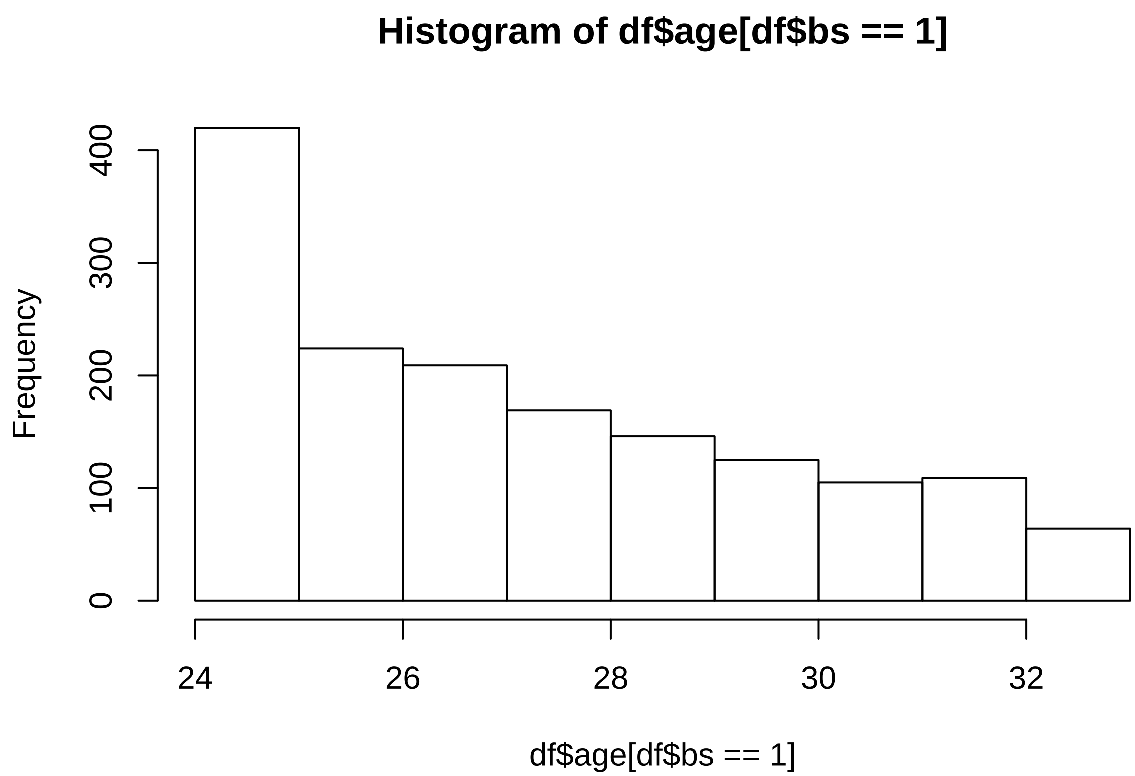
Table 2: Comparison of Age Missingness

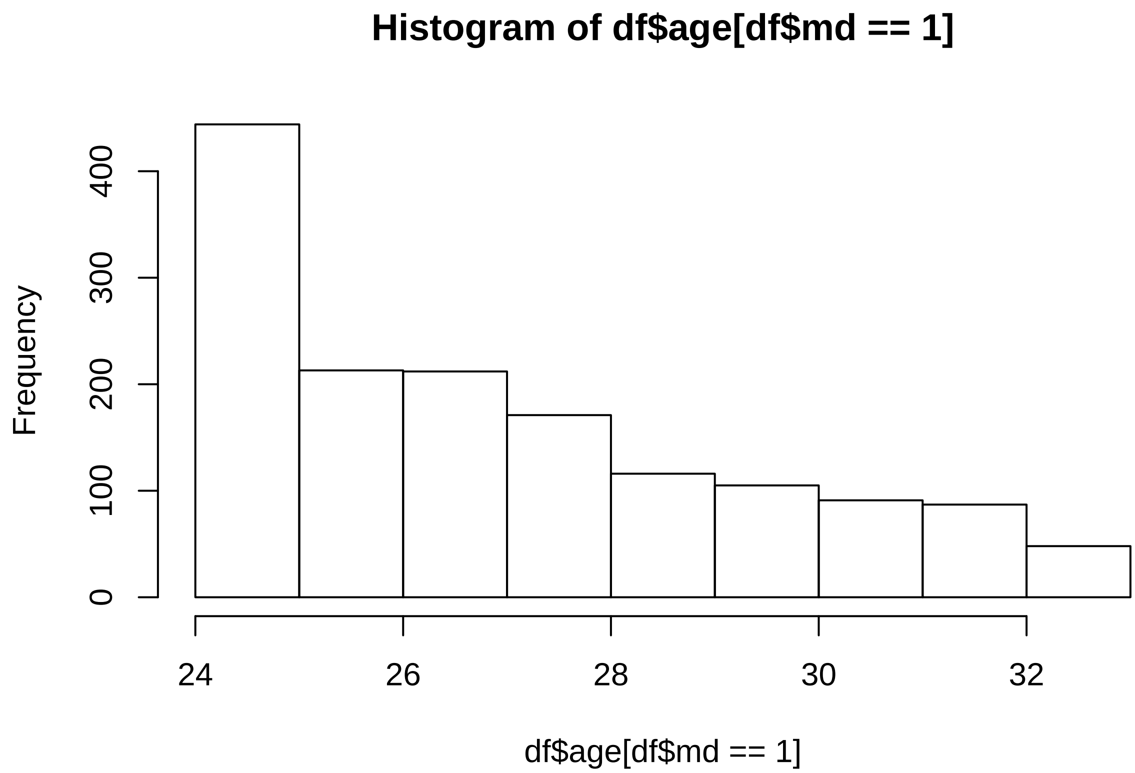
	Dependent variable:		
	Gender	Treatment	Location
	(1)	(2)	(3)
female	-0.014 (0.003)***		
bs		-0.002 (0.004)	
md		-0.001 (0.004)	
phd		0.004 (0.004)	
houston			-0.002 (0.005)
losangeles			-0.013 (0.006)*
newyork			-0.013 (0.006)*
philadelphia			0.004 (0.005)
phoenix			0.006 (0.004)
sanantonio			0.003 (0.005)
sandiego			-0.005 (0.006)
Constant	0.994 (0.001)***	0.987 (0.003)***	0.990 (0.004)***
Note:			*p<0.1; **p<0.05; ***p<0.01

EDA: Age Distribution

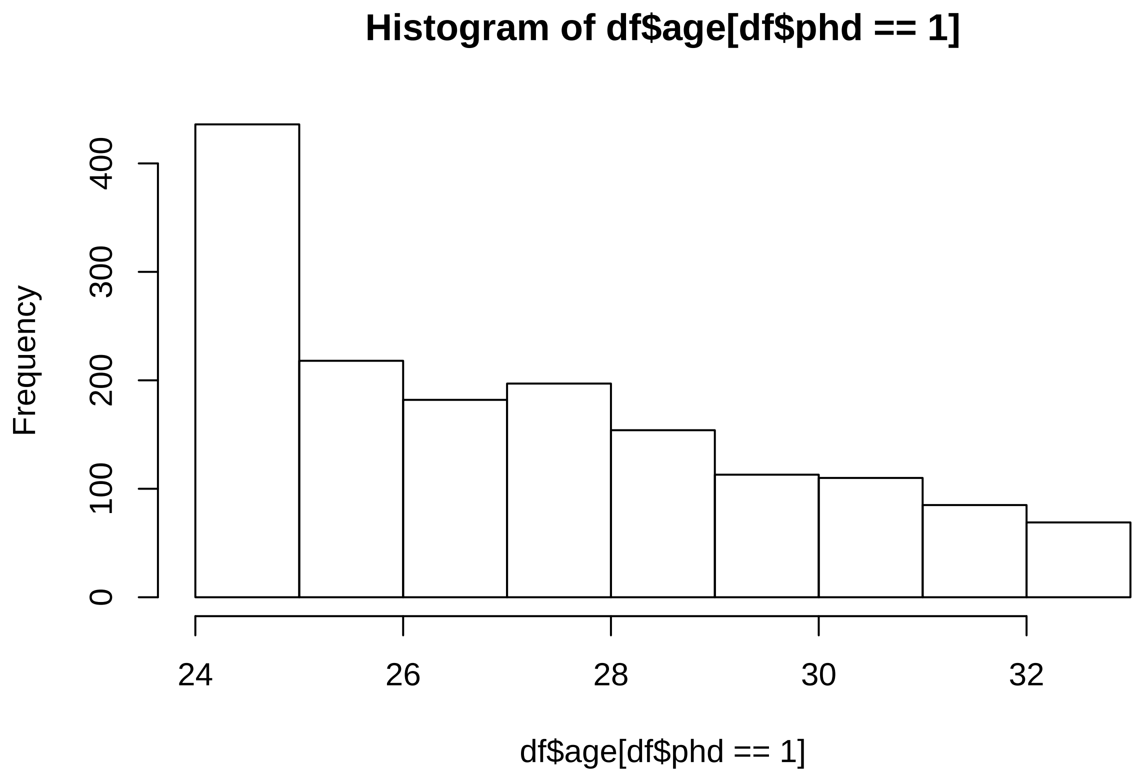
In addition to the missingness of covariates, the age distribution of suitors was also checked for balance across our treatment conditions. The test profiles had set an age filter of 24-34, and the suitor ages were expected to be evenly distributed in this range. However in every test condition and for every covariate, the suitors sampled had an age distribution that was centered near 24-25, with a strong right skew. Figure XXX below shows the ages of suitors in each treatment condition-no education shown, bachelor's degree, MD, and PhD.


Histogram of df$age[df$noedu == 1]
Frequency
0
100
200
300
400
24
26
28
30
32
df$age[df$noedu == 1]

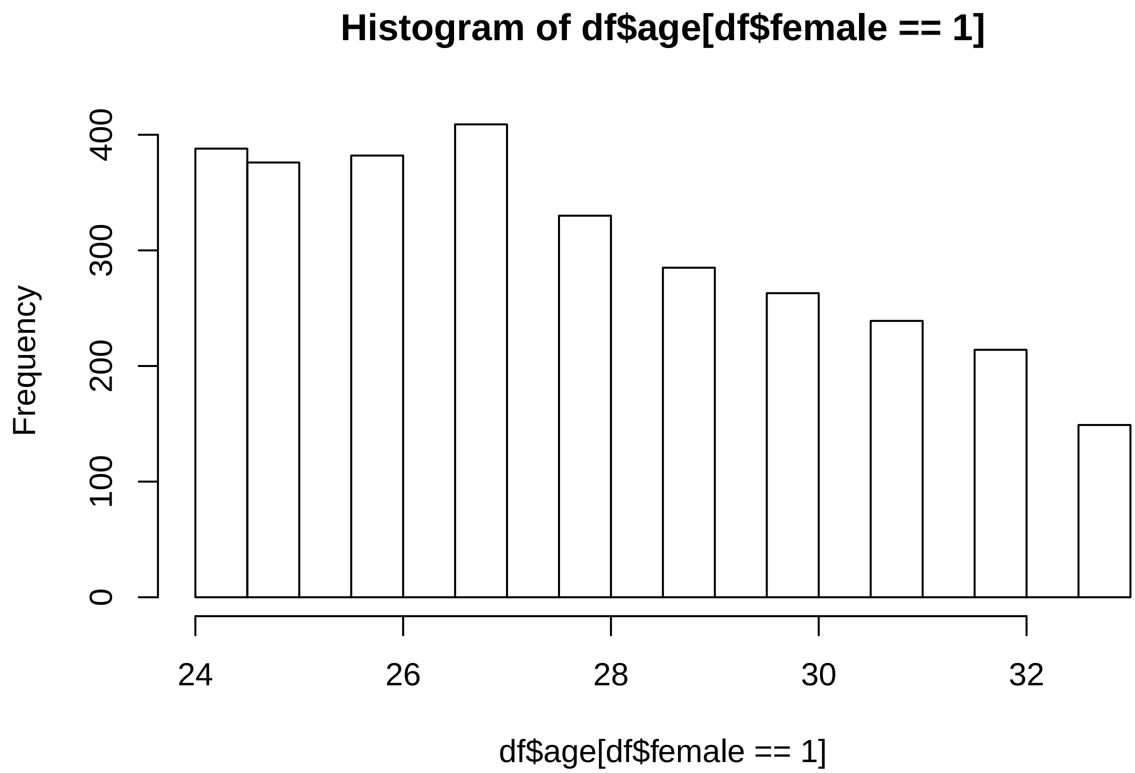

Histogram of df$age[df$bs == 1]
Frequency
0
100
200
300
400
24
26
28
30
32
df$age[df$bs == 1]

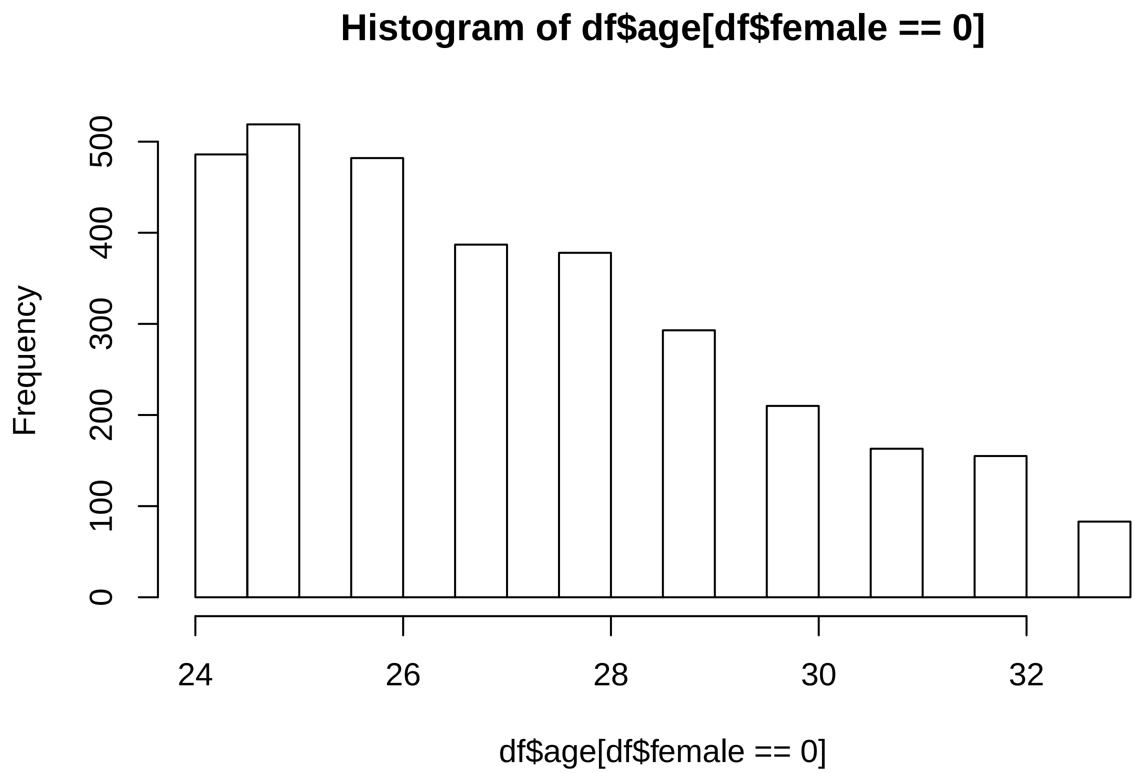


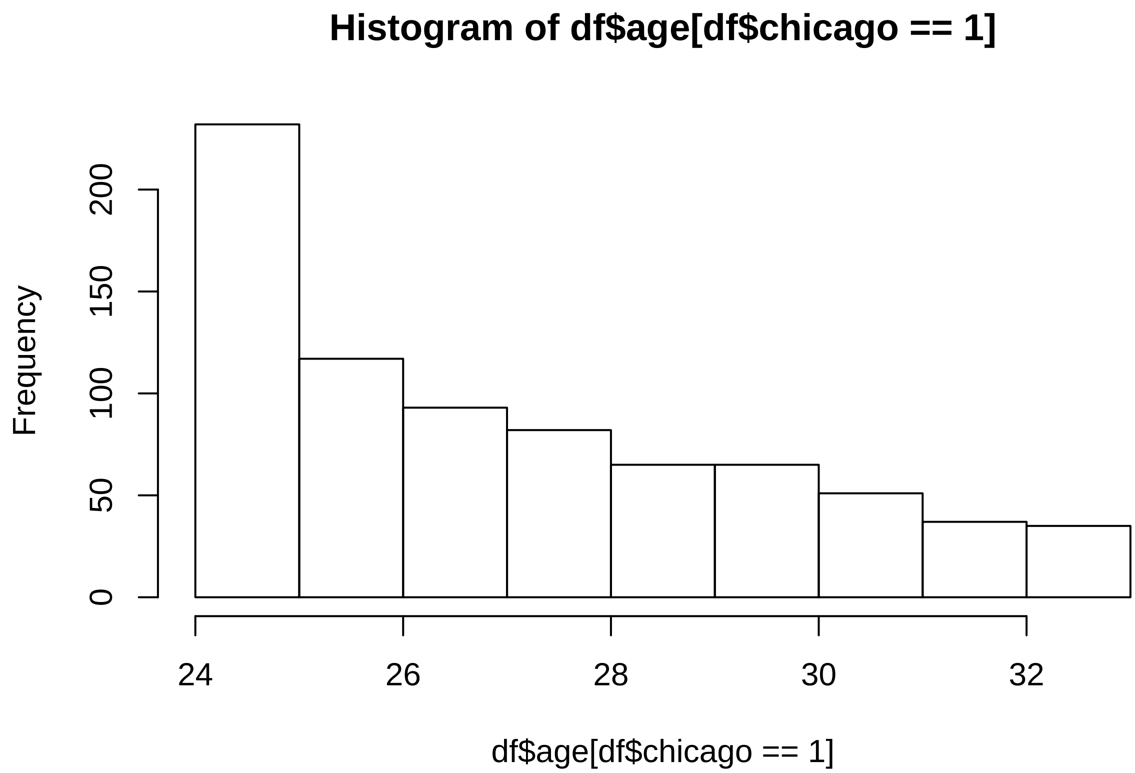
Histogram of df$age[df$md == 1]
Frequency
0
100
200
300
400
24
26
28
30
32
df$age[df$md == 1]

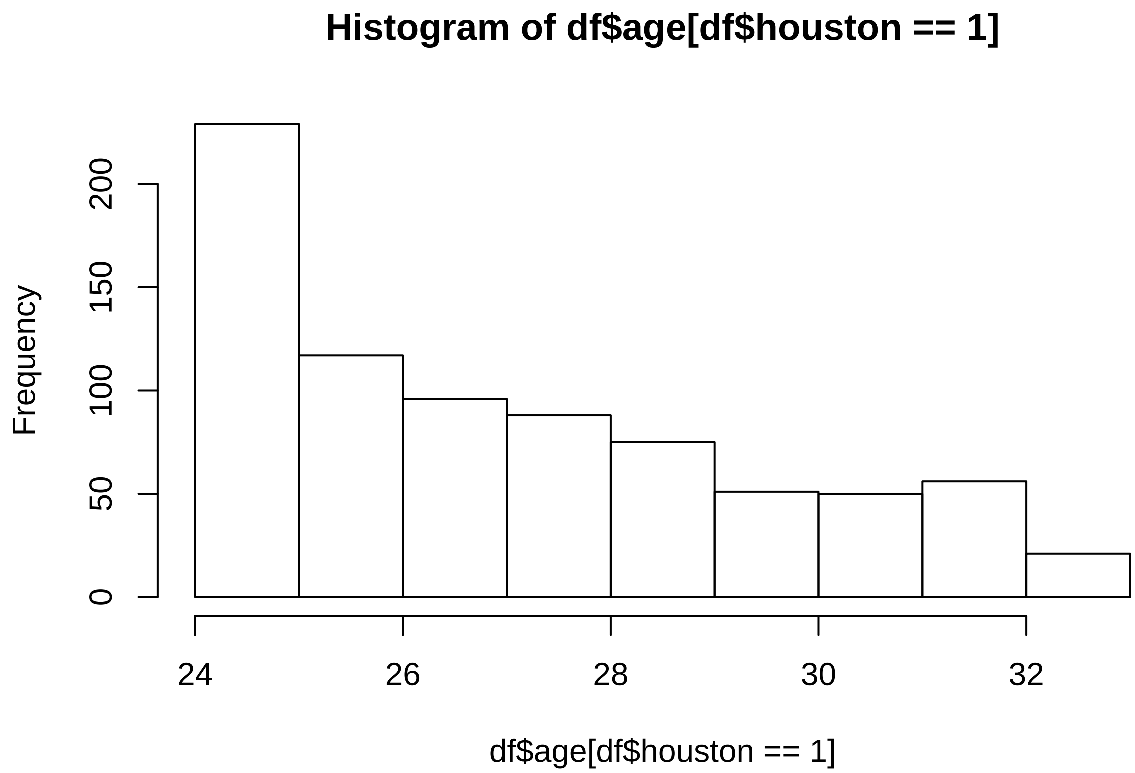


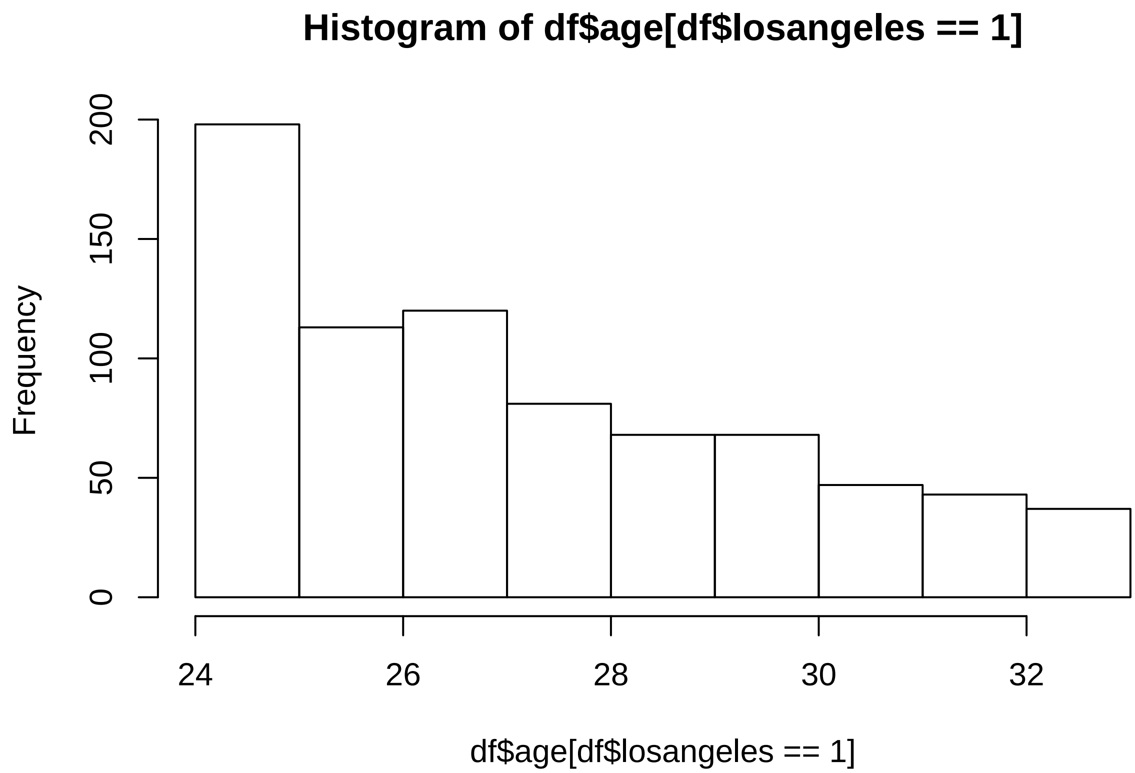
The strong skew in suitor age distribution raised a few questions about the use of Tinder to randomize treatment assignment– for example, does Tinder select suitors based on a profile's age or its age filter, or is it the case that the general population of Tinder users is around 25 years old? Figures XXX-XXX below show the ages of suitors for the other experimental conditions.

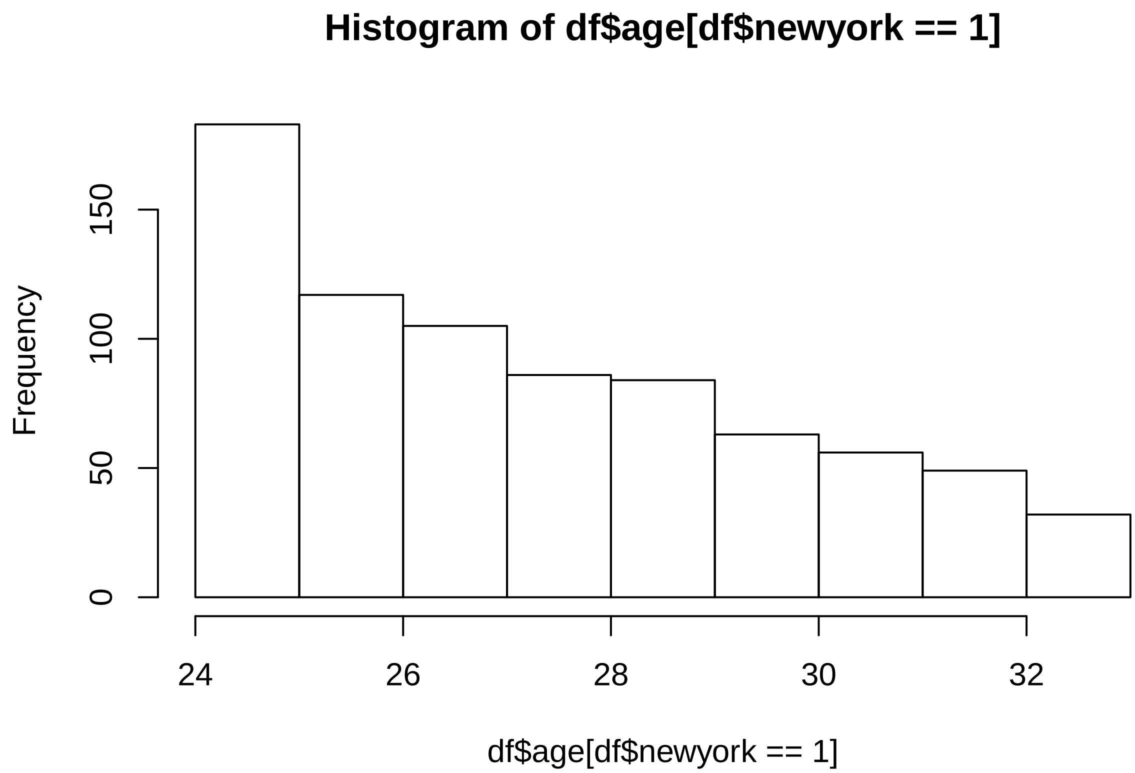

Histogram of df$age[df$female == 1]
Frequency
0
100
200
300
400
24
26
28
30
32
df$age[df$female == 1]

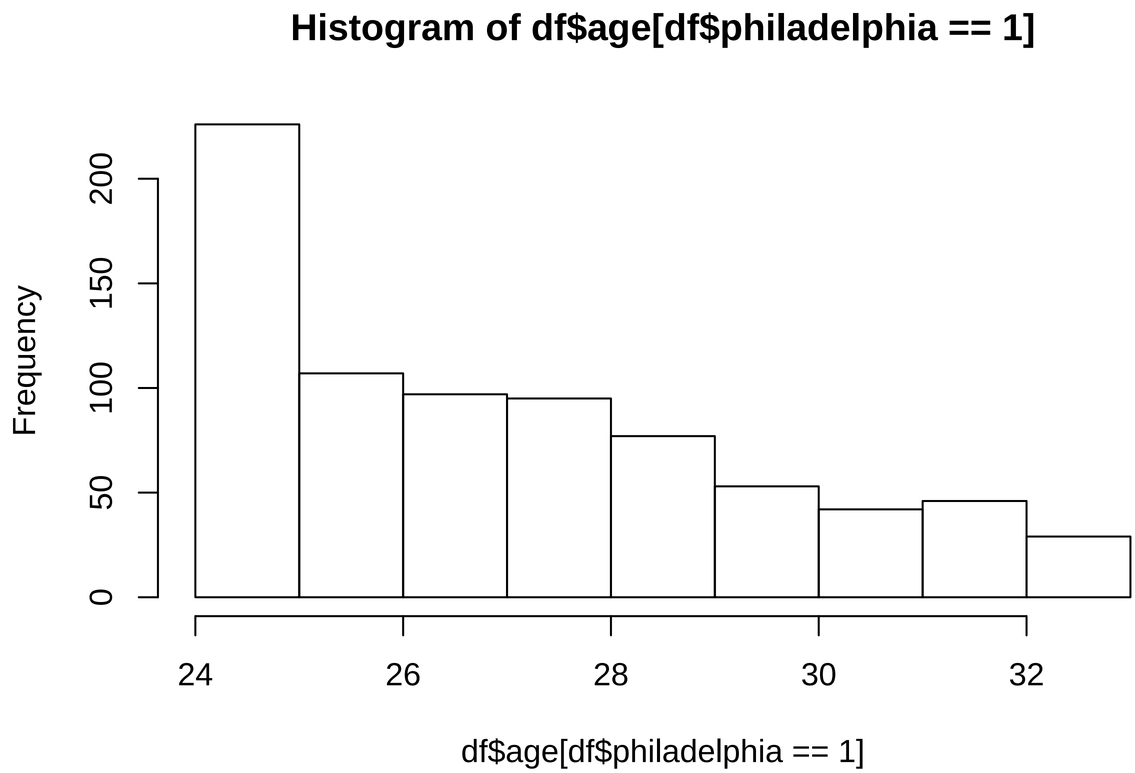

Histogram of df$age[df$female == 0]
Frequency
0
100
200
300
400
500
24
26
28
30
32
df$age[df$female == 0]

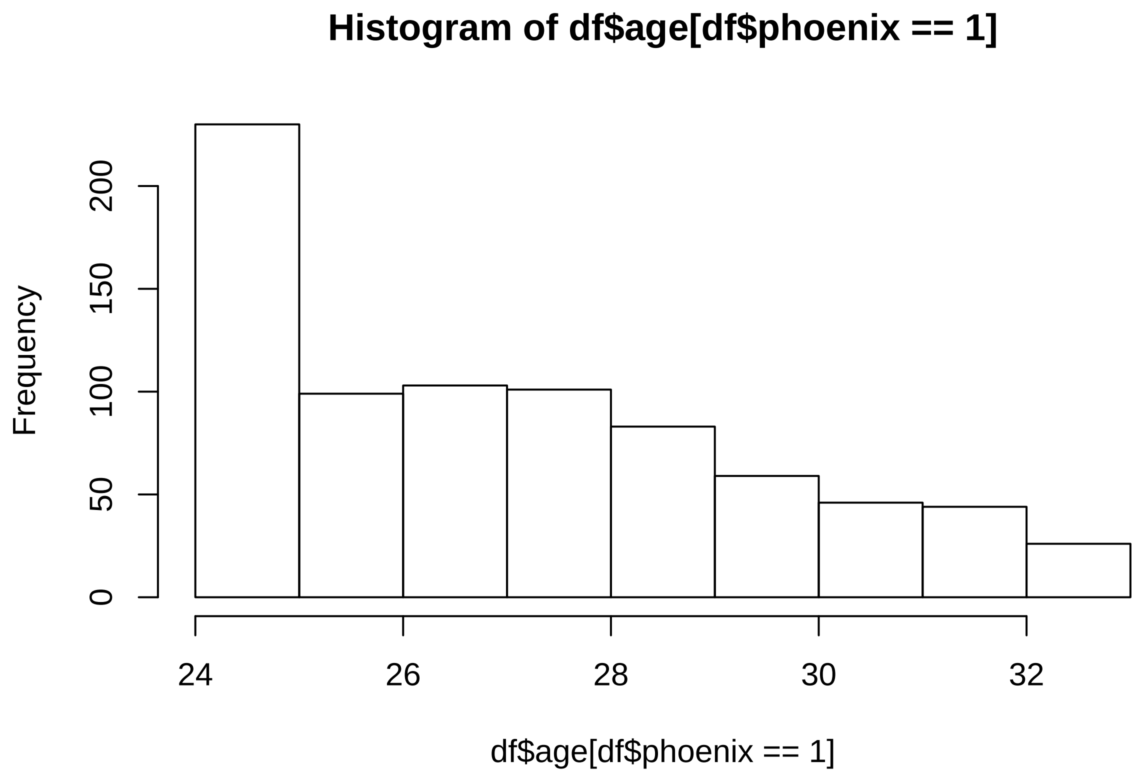

Histogram of df$age[df$chicago == 1]
Frequency
0
50
100
150
200
24
26
28
30
32
df$age[df$chicago == 1]

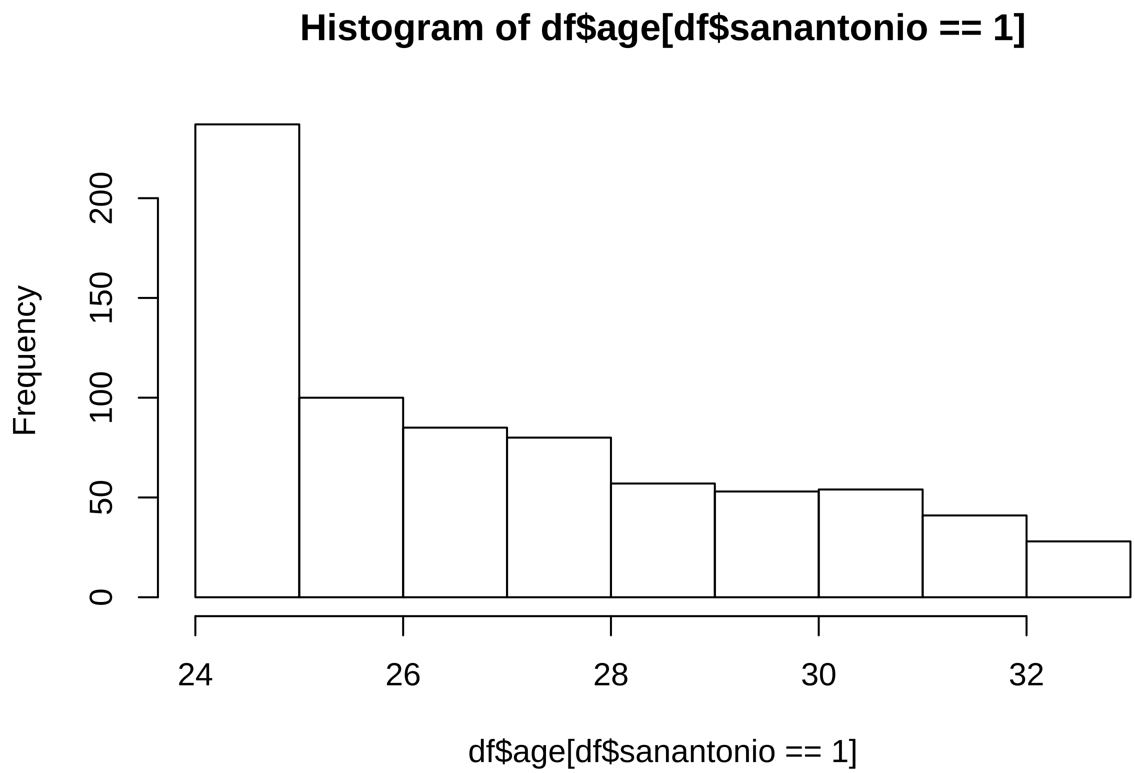

Histogram of df$age[df$houston == 1]
Frequency
0
50
100
150
200
24
26
28
30
32
df$age[df$houston == 1]

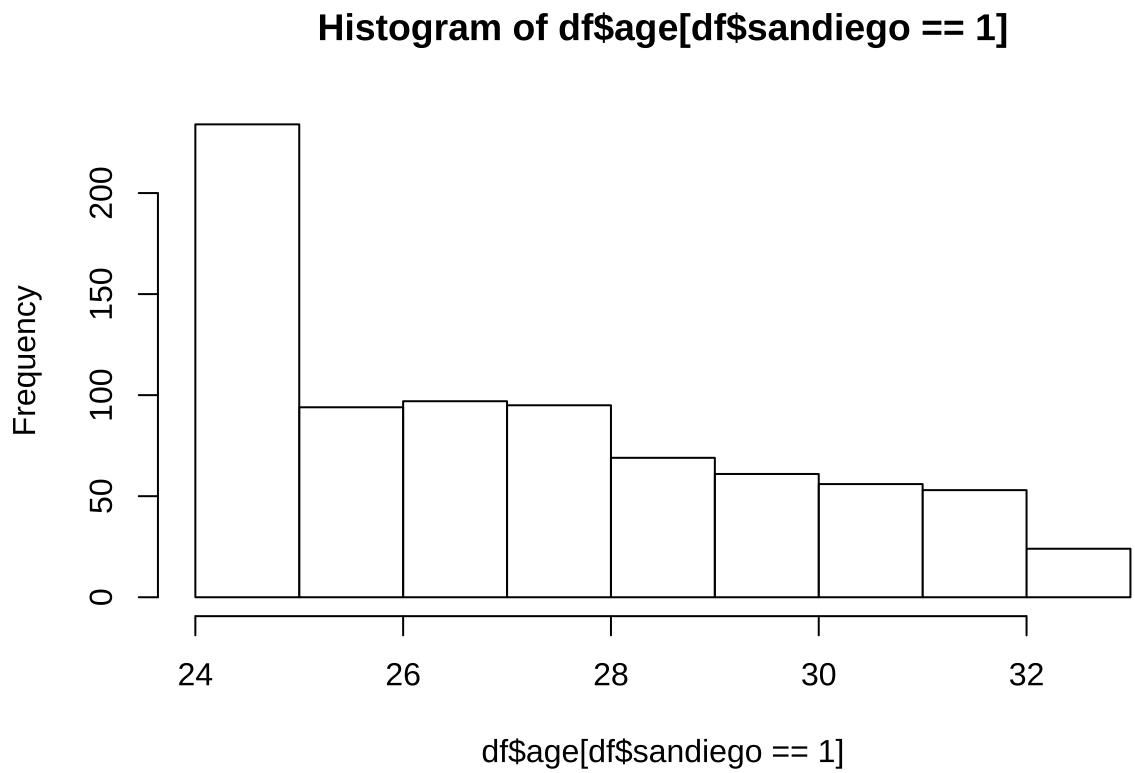

Histogram of df$age[df$losangeles == 1]
Frequency
0
50
100
150
200
24
26
28
30
32
df$age[df$losangeles == 1]


Histogram of df$age[df$newyork == 1]
Frequency
0
50
100
150
24
26
28
30
32
df$age[df$newyork == 1]


Histogram of df$age[df$philadelphia == 1]
Frequency
0
50
100
150
200
24
26
28
30
32
df$age[df$philadelphia == 1]


Histogram of df$age[df$phoenix == 1]
Frequency
0
50
100
150
200
24
26
28
30
32
df$age[df$phoenix == 1]


Histogram of df$age[df$sanantonio == 1]
Frequency
0
50
100
150
200
24
26
28
30
32
df$age[df$sanantonio == 1]



Fortunately, as shown in Figures XXX-XXX above, the skew in age distribution is very similar for all covariate values. Table XXX below shows the results of the tests for a difference in the average suitor age, by gender, treatment condition, and location. There appears to be a significant result that the male profile was shown suitors who were on average 0.59 years younger than those shown to the female profile, but for this experiment, a half-year is considered a small effect size. Most importantly, no significant difference in average suitor age was found between treatments, so the experiment still yields an apples-to-apples comparison.

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 16, 2017 - 10:05:11 PM

Table 3: Comparison of Average Age

	Dependent variable:		
	Gender	Treatment	Location
	(1)	(2)	(3)
female	0.590 (0.066)***		
bs		0.171 (8.922)	
md		$-$0.091 (0.000)***	
phd		0.099 (0.000)***	
houston			$-$0.012 (0.132)
losangeles			0.143 (0.133)
newyork			0.280 (0.132)**
philadelphia			$-$0.009 (0.133)
phoenix			0.005 (0.131)
sanantonio			$-$0.072 (0.137)
sandiego			0.061 (0.134)
Constant	27.171 (0.045)***	27.414 (609.994)	27.411 (0.094)***
Note:			*p<0.1; **p<0.05; ***p<0.01