Final Report on How Couples Meet and Stay Together (HCMST) Dataset

Dawn Chen, James Diao, Jørn Emborg, Amy Zhao

May 4, 2017

Contents

1 Introduction **1**

2 Data **2**
2.1 Variable Descriptions 2
2.2 Exploratory Plots 3

3 Analysis **5**
3.1 Methods of Meeting and the Rise of Online Dating 5
3.2 The Demographics of Online Meeting 6
3.3 The Role of Religion and Race in Partner Selection 7
3.4 Nota Bene 10

4 Conclusion **10**

5 Code Appendix **11**

1 Introduction

Romantic relationships are an area of broad interest. Love, dating, marriage, and divorce constitute some of the most significant events in many people's lives. With emerging data, it may be possible to better understand these important interpersonal interactions, and how they have changed over time.

Michael Rosenfeld's group at Stanford has compiled a dataset entitled "How Couples Meet and Stay Together" (HCMST), containing 534 survey responses from 4002 respondents on demographic and relationship information (https://data.stanford.edu/hcmst).

HCMST contains 5 sequential rounds of surveys from 2009, 2010, 2011, 2013, and 2015. Due to response bias in follow-up surveys, we have refrained from speculating on which couples have stayed together and why. Instead, we have focused on the earlier datasets, aiming to draw conclusions from the couples as they reported themselves in 2009.

Using this data, we hope to examine the factors associated with relationship satisfaction and longevity, including:

1. How has the way couples meet changed over time?
2. What factors predict whether a respondent met their partner online?
3. How do people select partners based on religion and race?

2 Data

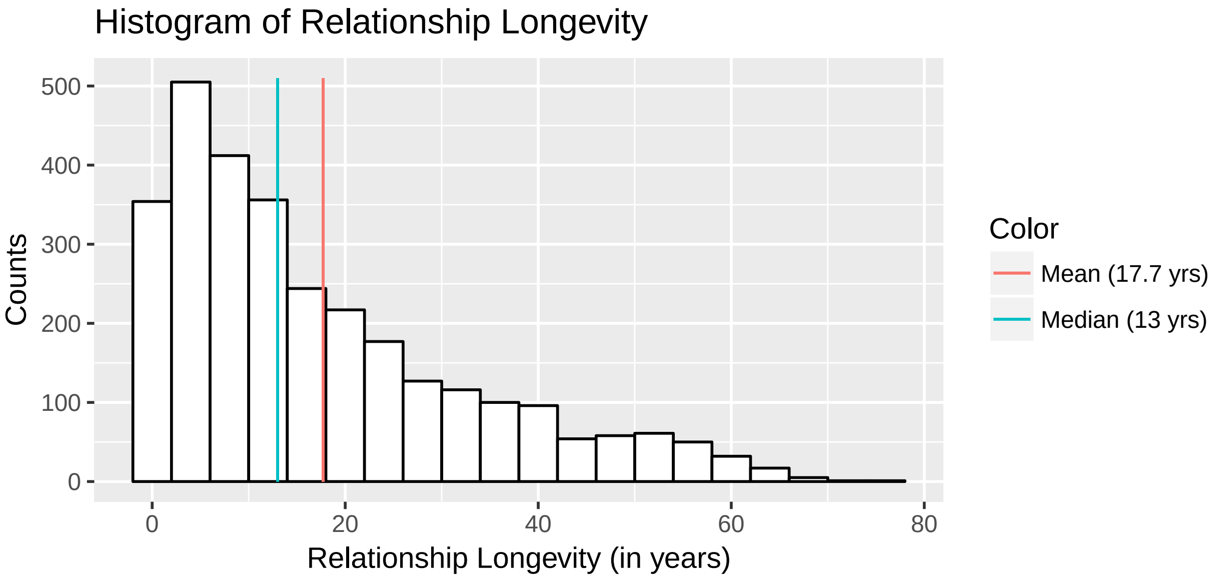
2.1 Variable Descriptions

1. PPAGE: respondent age at time of HCMST wave I survey (from 19-95); integer: range from 19-95
2. PPGENDER: respondent gender; binary: range from 1-2 (1 = male, 2 = female)
3. CHILDREN_IN_HH: number of children in household; integer: range from 0-7
4. PPMARIT: marital status; categorical factor: married, widowed, divorced, separated, never married, living with a partner
5. HHINC: dollar value household income; integer: range from 2500-200000
6. PPWORK: current employment status; categorical factor: working – as a paid employee, working – self-employed, not working – on temporary layoff from a job, not working – looking for work, not working – retired, not working – disabled, not working – other
7. PPEDUCAT: educational status of respondent; categorical factor: less than high school, high school, some college, bachelor's degree or higher
8. PAPRELIGION: identified religion of respondent; categorical factor: Baptist, Protestant, Catholic, Mormon, Jewish, Muslim, Hindu, Buddhist, Pentecostal, Eastern Orthodox, other Christian, other non-Christian, None
9. PPPARTYID3: political party affiliation; categorical factor: republican, democrat, independent, another party, no preference
10. RESPONDENT_RACE: race that the respondent identifies with; categorical factor: Non-Hispanic White, Non-Hispanic Black, Non-Hispanic American Indian, Non-Hispanic Asian/Pacific Islander, Non-Hispanic Other, Hispanic
11. PARTNER_RACE: race that the respondent's partner identifies with; categorical factor: Non-Hispanic White, Non-Hispanic Black, Non-Hispanic American Indian, Non-Hispanic Asian/Pacific Islander, Non-Hispanic Other, Hispanic
12. RESPONDENT_YRSED: educational attainment of respondent; numeric: range from 0-20
13. PARTNER_YRSED: educational attainment of respondent's partner; numeric: range frmo 0-20
14. W4_ATTRACTIVE: respondent's rated attractiveness of themselves; integer: range from 1-4 (4 = very attractive, 3 = moderately attractive, 2 = slightly attractive, 1 = not at all attractive)
15. W4_ATTRACTIVE_PARTNER: respondent's rated attractiveness of their partner; integer: range from 1-4 (4 = very attractive, 3 = moderately attractive, 2 = slightly attractive, 1 = not at all attractive)
16. HOW_LONG_RELATIONSHIP: current relationship duration; integer: range from 0-76
17. GENDER_ATTRACTION: sexual preference; categorical factor: opposite gender only, mostly opposite, both genders equally, same gender mostly, only same gender
18. SAME_SEX_COUPLE: sexual preference; binary: 0 = opposite-sex couple, 1 = same-sex couple
19. PARENTAL_APPROVAL: parental approval or disapproval of respondent's relationship with partner; binary: 0 = disapproval or unknown, 1 = approval
20. RELATIONSHIP_QUALITY: respondent's assessment of relationship quality with partner; integer: range from 1-5 (1 = very poor, 2 = poor, 3 = fair, 4 = good, 5 = excellent)
21. HOW_LONG_RELATIONSHIP: length of respondent's most recent relationship; integer: range from 0-76
22. AGE_DIFFERENCE: absolute difference in age between the respondent and their partner; integer: range from 0-69
23. RESPONDENT_RELIG_16_CAT: respondent's identified religion at 16; categorical factor: Protestant or oth Christian, Catholic, Jewish, Neither, No religion
24. PARTNER_RELIG_16_CAT: partner's identified religion at 16; categorical factor: Protestant or oth Christian, Catholic, Jewish, Neither, No religion
25. HOW_LONG_AGO_FIRST_MET: time since respondent first met partner; numeric: 0-76
26. HOW_LONG_AGO_FIRST_ROMANTIC: time since respondent was first romantically involved with partner; numeric: 0-76
27. MET_?: whether the respondent met their partner in the specified fashion; binary: 1 = Yes, 0 = No
28. YEAR_MET: indicates the year that the respondent first met their partner; numeric: 1933-2009

2.2 Exploratory Plots

2.2.1 Histogram of Relationship Longevity

How long do people typically stay together? This question is answered by the distribution of relationship lengths.

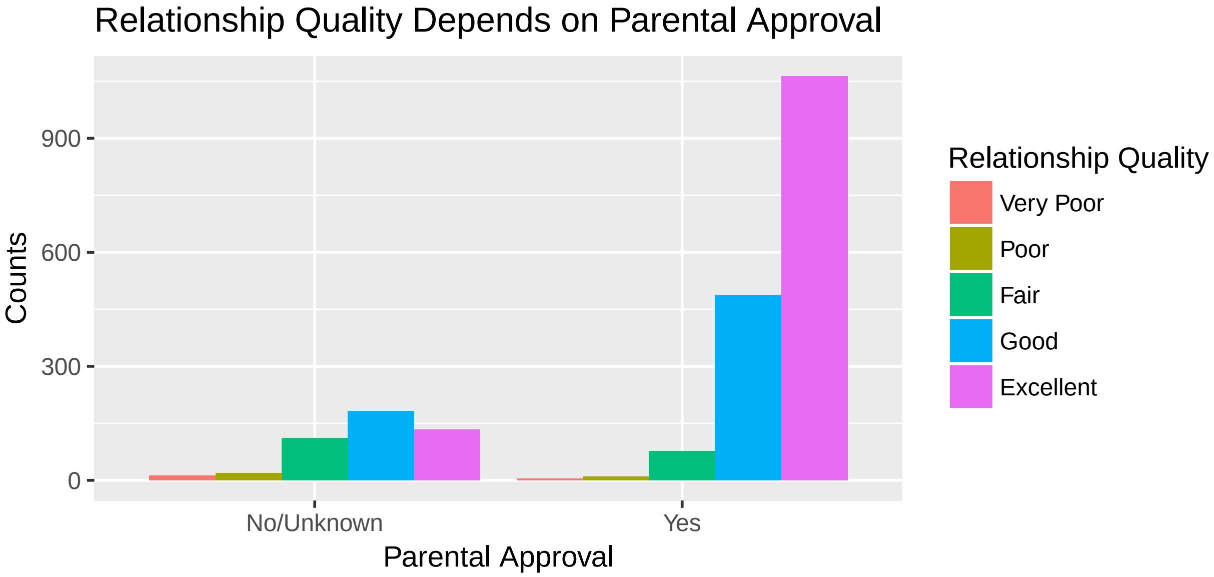


Most relationships have been short-term, leading to a right-skew in the histogram. The relationship length in years for increments of 10% of respondents are given by:

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90% 100%
##   0.0   1.5   4.0   7.0  10.0  13.0  17.0  22.4  30.0  42.0 76.0
```

2.2.2 Parental Approval and Relationship Quality

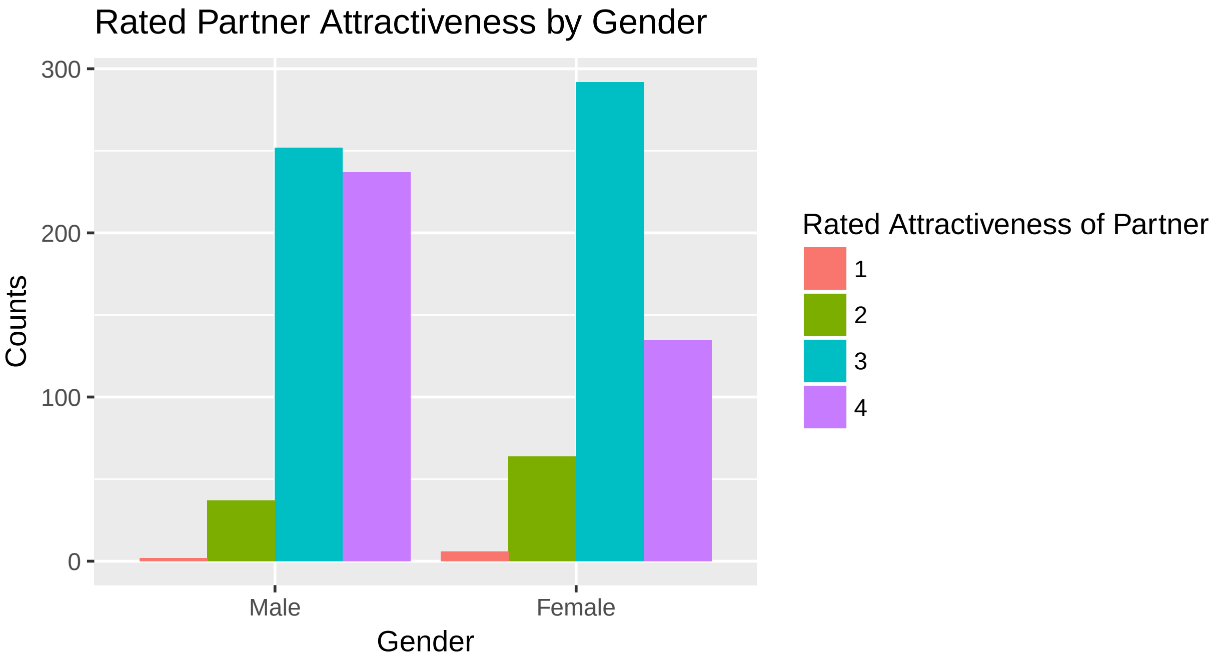
A preliminary linear model showed that parental approval was the strongest linear predictor of relationship quality. This can be visualized as split distributions.



Couples with parental approval are much more likely to rate their relationship quality as “excellent” (65%), compared to those without or with unknown parental approval (29%). In fact, those without or with unknown parental approval report a higher percentage of “good” relationships than “excellent” relationships. This result is interesting but requires a better understanding of confounding variables to verify.

2.2.3 Importance of Attractiveness by Gender

We were interested to find that the attractiveness rating of the partner was a survey question. We aimed to stratify this by various demographics, and found that the most robust difference was by gender.



It is clear that men are much more likely to rate their partners as more attractive, and women are much more likely to rate their partners as less attractive. In fact, men are 66% more likely to rate their partners as “very attractive”, while women are 3 times more likely to rate their partners as “not attractive at all.” This suggests that men are much more likely to value attractiveness in a partner as see this as a precondition for a relationship.

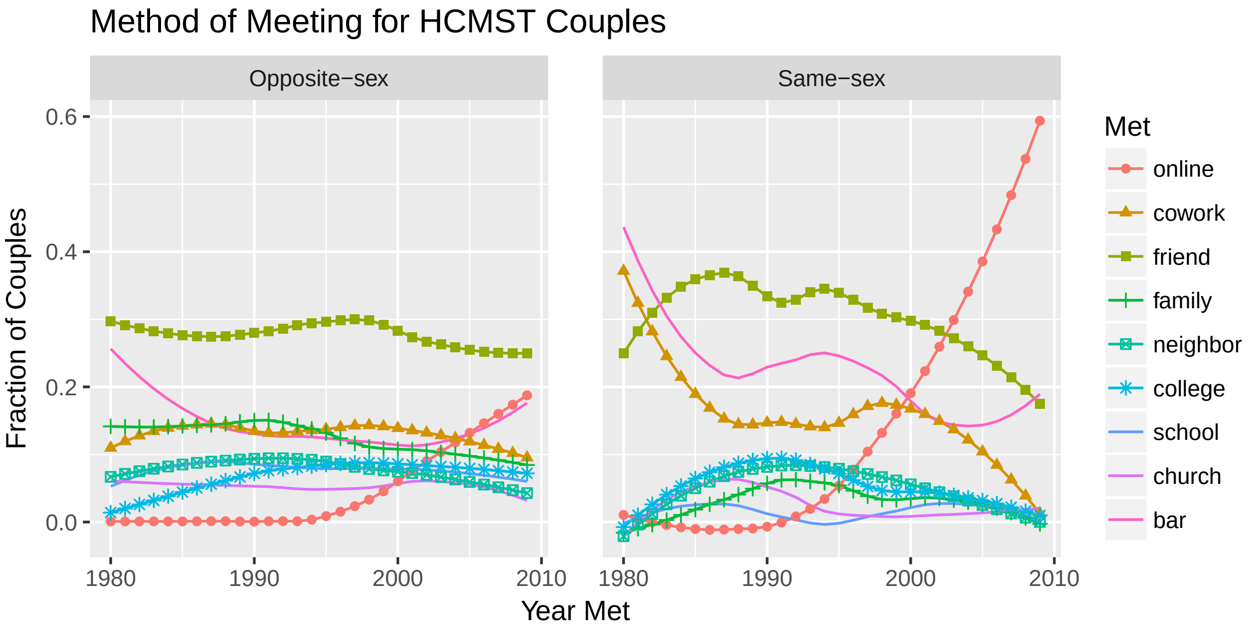
3 Analysis

3.1 Methods of Meeting and the Rise of Online Dating

There are many ways in which couples meet, including through family, friends, the church or the neighborhood. With the rise of the Internet, communications within existing social networks have become more prevalent, and it has become much easier to search for and find potential partners outside of one's immediate social network. Indeed, the number of users of online dating platforms - such as Tinder - has been on the rise.

The HCMST dataset contains data from couples who met at diverse time points. We decided to analyze how the ways in which couples meet have changed over time. This is based on their response to the question "How did you meet your partner?", and the year they met was calculated from the variable HOW_LONG_AGO_FIRST_ROMANTIC. The meeting trends for both same-sex and opposite-sex couples from 1980 to 2009 (smoothed with LOWESS) are displayed below.

Note: N = 2462 for opposite-sex couples, and N = 462 for same-sex couples. These were coded as 1 and 2 in the SAME_SEX_COUPLE variable, respectively.



For couples who met in 1990 or before, essentially no one met online. This makes sense, because the Internet had not yet been popularized. Between 1995 and 2005 for opposite sex couples and between 1990 and 2010 for same-sex couples, there is an approximately linear growth in the proportion of respondents who met their partners online. By 2009, 20% of opposite-sex couples had met online. At the same time, more than 60% of same-sex couples met online. We can see that online meeting has become the dominant and preferred method of meeting for same-sex couples. There are multiple possible explanations. One is that online dating expands options more for those who have fewer options in the first place (i.e., gay and lesbian individuals). Alternatively, this effect could be driven by the increased anonymity online, where individuals are less likely to suffer shame or social consequences for being gay/lesbian.

For opposite-sex couples, off-line ways of meeting have remained relatively steady from 1980 to today. Meeting romantic partners through friends is true of the plurality of our respondents, hovering at about 30%. Meeting through school, family, neighbor, church or coworkers have also remained constant. There is a slight drop in the percentage of people who meet at bars, perhaps because the internet has replaced going to bars as a new way of meeting strangers.

For same-sex couples, there is much more variability in the trends, which may be attributed to the smaller sample size.

3.2 The Demographics of Online Meeting

In the previous section, we noted a dramatic increase in the number of couples that met online, especially for same-sex couples. This finding motivated us to further explore relationships between meeting online and other variables, such as religion, political party, and education level. Specifically, we were interested in the distribution of percentages of couples of different religions, political parties, and education levels who either (1) met online or (0) did not meet online. We could visualize this "distribution" of different percentages through the use of tables.

Let us first look at the distribution of couples who met online or offline among different political parties.

	(1) republican	(2) other	(3) democrat
Met Offline	0.41	0.024	0.57
Met Online	0.31	0.011	0.68

From this table, it looks like more of those who meet online are democrats, as opposed to those who meet offline (67.8% vs. 57.0%, respectively). Using Pearson's chi-squared test - which is a statistical test that looks at how likely observed difference between categorical variables occurs by chance, we find that this difference among democrats, republicans, and other political parties in meeting online and offline is significant. We computed a p-value of 0.002226, which is significant at a significance level of 0.05, even after correcting for multiple testing. This finding may be due to the fact that democrats are younger and more tech-savvy, and thus more open to online dating.

Next, we investigated the relationship between meeting online and the respondent's education level:

	Less than high school	High school	Some college	Bachelor's degree
Met Offline	0.110	0.26	0.28	0.35
Met Online	0.037	0.14	0.39	0.43

Similarly, there highly educated people are better represented in the couples that meet online as compared to those who meet offline (p-value for Pearson's Chi-Squared test of 1.58×10^{-8}). In fact, there is an increase in the fraction of people who are highly educated (at least some college) who meet online (39% and 43% respectively) compared to those who meet offline (28% and 35%). At the same time, there is a decrease in representation of people with lower levels of education in the group of couples that meet online; the fraction of people with a high school education or less, decreases from 11% 26% for level 2 in the offline group to 4% and 14% in the online group, respectively. This finding may be a result of access to technology. For example, the more educated someone is, the more likely they may be able to afford computers and smartphones, which consequently allow for easy access to online dating sites.

Finally, let us look at the correspondence between parental approval and meeting online:

	Disapproval or Unknown	Approval
Met Offline	0.20	0.80
Met Online	0.32	0.68

The difference among groups here is again statistically significant; Pearson's Chi-Squared test yields a p-value of 1.58×10^{-8}. Of the people who meet online, there is a decrease in those with parental approval as compared to people who meet offline. This may be correlated with younger age or other factors.

We looked at associations between online dating and other variables (with the exception of religion, which will be discussed in the next section), but did not find any statistically significant associations.

3.3 The Role of Religion and Race in Partner Selection

More online dating services have been created targeting specific religious groups (for example, Christian Mingle) than ever before. We were interested in whether this trend is on to something - do people tend to find couples within the same religious group? We also examined racial tendencies; while it is politically incorrect to target dating sites for specific racial/ethnic groups, the same trend to find potential partners within the same race may be statistically significant.

In this section, we will investigate whether couples tend to adhere to the same religious and racial groups when choosing partners.

3.3.1 Respondent Filtering

As there was no variable that explicitly stated whether the respondent was in a relationship or not (the PPMARIT variable states marital status, but couples can be in a relationship and not necessarily married), we estimated people who were not in relationships as those respondents with missing values (i.e. NAs) in the partner columns. For example, only 3% of values were missing for the respondents while about 25% of values were consistently missing from the variables corresponding to partners. Therefore, we claim that these missing values indicate which respondents are single. Using this conservative method, we selected for people who had or currently have partners with information on race and religion, which we will investigate further in the following sections.

3.3.2 The Role of Religion in Partner Selection

To explore whether respondents tend to end up with partners that are of the same religion, we created a binary variable, RELIGION_SAME that has value 0 if the couple does not believe in the same religion (PARTNER_RELIG_16_CAT != RESPONDENT_RELIG_16_CAT) or 1 if they do indeed believe in the same religion (PARTNER_RELIG_16_CAT == RESPONDENT_RELIG_16_CAT). Note that if both respondent and partner are non-religious, this variable still counts them as being of the same religion. Furthermore, it is important to note that the 'Neither Christian nor Jewish' category may pose problems since it includes various other religions, such as Islam, Hindu, and Buddhism. However, we found that the number of people in these categories was relatively small (in fact, there were a total of 13 matches between partners with religions in this category out of 2991 respondents), and unlikely to affect the overall findings.

The fraction of couples with the same religion is computed to be: 56.6%

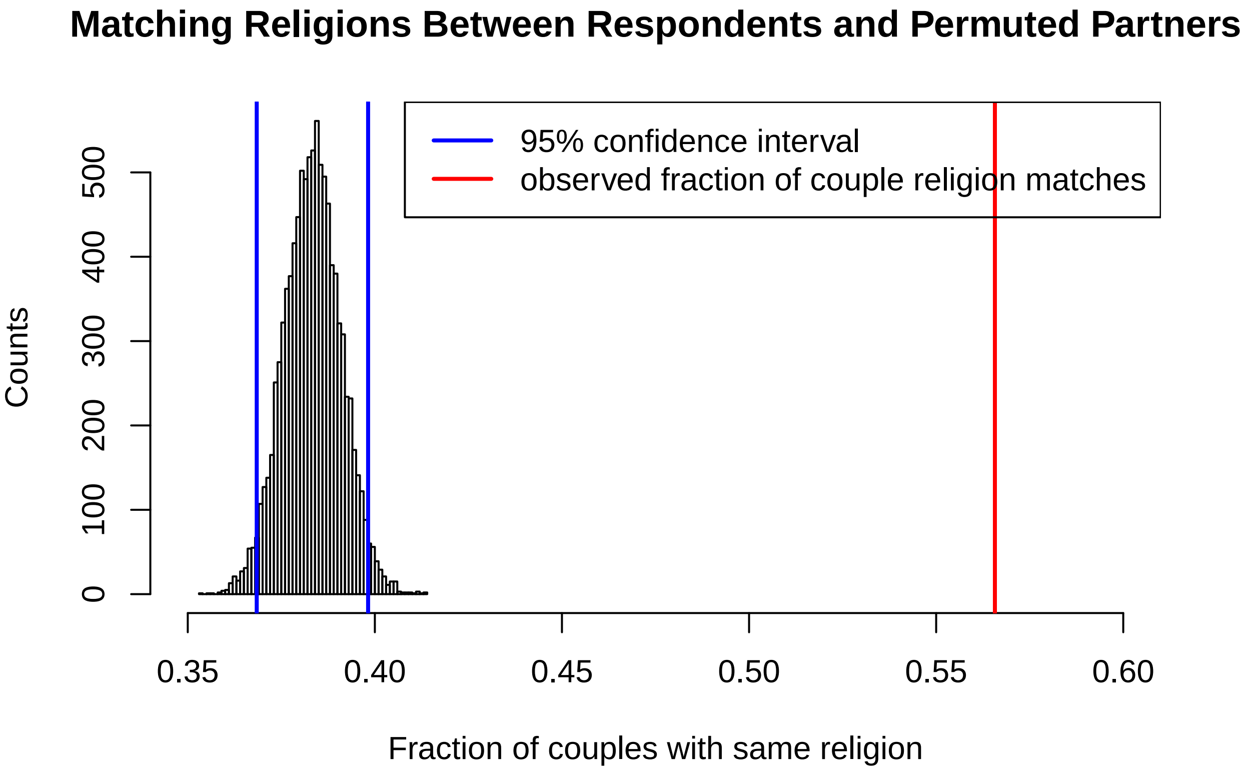
That is, there are 56.6% of respondents that have the same religion as their partners. However, this number does not tell us much on its own. To make sense of this percetange, we can compare it to what we would expect in a sample in which the partners' religions are matched at random.

In order to investigate this, we run a permutation test, where we shuffle up the respondents' partners and find the new proportion of same-religion couples.

The mean fraction of matches between respondent and permuted partner religions is approximately 38.3%.

How significant is this difference? The 95% confidence interval and histogram from the permutation test are shown below:

2.5%	97.5%
0.3684303	0.3981946



This graph shows that the observed fraction of couples with the same religion is statistically significant. Thus, people almost certainly consider religion as a strong determining factor in choosing their significant others.

3.3.3 Online vs. Offline Selection of Same Religious Partners

Lastly, we wanted to see if there is an effect on the representations of different religious groups from targeting online dating services. Using a similar Chi-Square test from the previous section, we have:

	Different Religion	Same Religion
Met Offline	0.43	0.57
Met Online	0.54	0.46

Interestingly, there is a statistically significant decrease in the fraction of couples with the same religion who meet online as opposed to those who meet in other methods (Pearson Chi-Squared test with Yates' continuity correction with 1 degree of freedom yields a p-value of 0.0007). This effect may be due to the stigmatized use of same religion dating sites (as using these sites make people conciously aware of their choice to date within their in-group, while real-world dating may not include this conscious appraisal of the partners' religion).

3.3.4 The Role of Race in Partner Selection

Next, we delve into an exploration of how race affects partner selection. That is, do most people marry someone of the same race or someone of a different race?

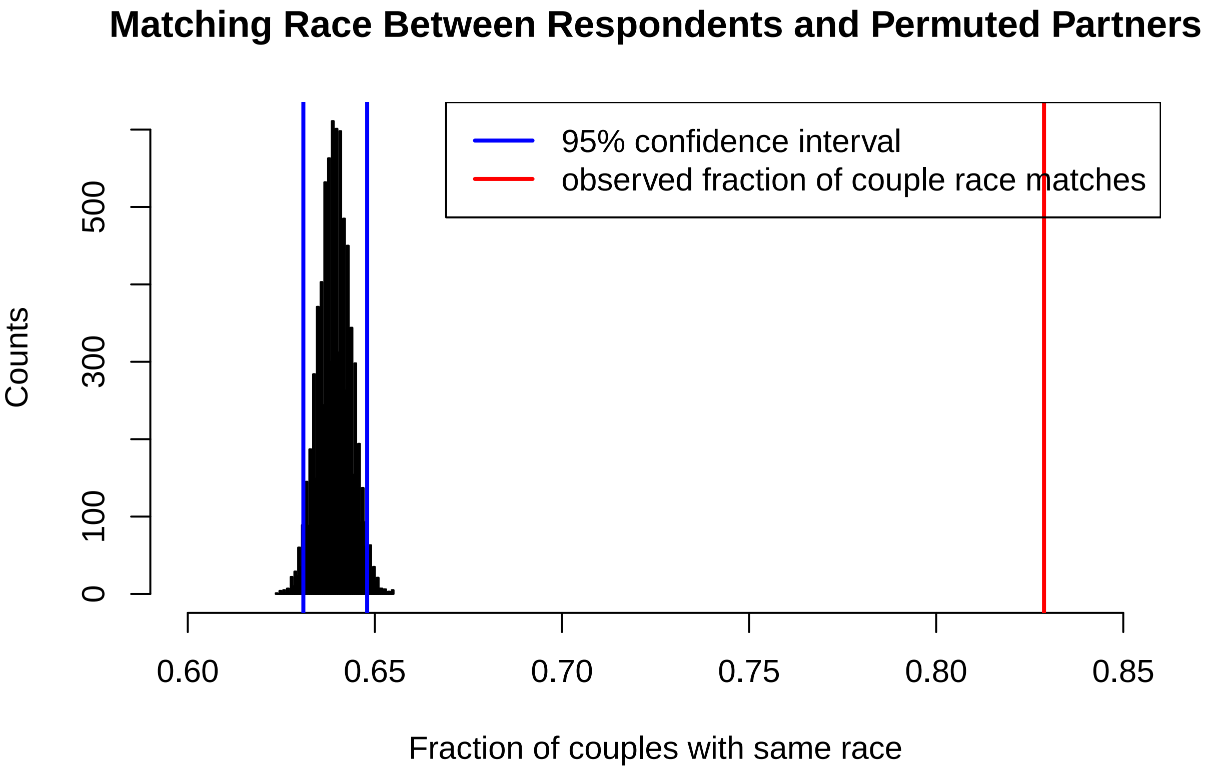
Similar to our analysis of religion, we create a binary variable, RACE_SAME, with 1 indicating that the respondent and partner are of the same race (RESPONDENT_RACE == PARTNER_RACE) and 0 indicating that the couple is of different races (RESPONDENT_RACE != PARTNER_RACE). The percentage of respondents who are the same race as their partner is computed as: 82.9%.

That is, there are 82.9% of respondents are of the same race as their partners. While this number may seem high, it is not sufficient to conclude statistically significance. In order to do so, we must run another permutation test (that is, where the partners' races are randomly sampled without replacement).

From this simulation, the mean of percentage of matches between respondent races and permuted partner races is approximately 63.9%.

The corresponding 95% confidence interval and histogram are shown below:

2.5%	97.5%
0.6308927	0.6479438



Again, it is obvious that the observed fraction of couples with the same race is statistically significant as the observed value is far from the simulated distribution of fraction of matches, which suggests that race plays an important role in partner selection. That is, on the population level, people tend to strongly prefer partners of the same race.

We conducted a further analysis of how much race plays a role within different racial groups. We found that 92.6% of white people, 76.7% of African Americans, 16.7% of Native Americans, 39.3% of Asian Americans, and 37.3% of Hispanics have partners of the same race. Now, the last three races seem to have a surprisingly low number of people who have partners of the same race; however, this observation can be accounted for by the low number of respondents who are of these races (only representing 1, 2, and 10 percent of the total respondents, respectively).

This result suggests that people do indeed marry within their same race, with this effect the most stronlgy observed in white couples.

3.4 Nota Bene

With a large dataset like HCMST, if we look hard enough, we will find statistically significant results at a given level. That is, if we run 100 independent tests, 5 tests will show a significance level less than 0.05. To account for this, we may use the Bonferroni correction (divide the statistical significance level by the number of tests conducted). For example, in this dataset, there may be up to 20 features to consider for their associations with online dating. We then would divide 0.05 by 20 to get an adjusted significance level of 0.0025. However, for our purposes, this correction did not change our conclusions about statistical significance.

4 Conclusion

The HCMST has given us the opportunity to take common understandings of dating and relationships, and justify them empirically.

First, we found that more recent relationships are much more likely to have met online. This is by far the most robust change in the way couples meet, and particularly striking in same-sex couples.

Second, we found that meeting online is not uniform across demographic groups. Individuals who identify as Democrats, who are more highly educated, and who lack parental approval, are more likely to have met online.

Third, we investigated the role of religion and race in partner selection, finding that individuals strongly prefer individuals who are similar to them. The findings are significant even after correcting for multiple hypothesis testing.

Our primary concerns with the data are with possible selection bias. Although phone banking is one of the more reliable ways of collecting a representative sample, it will tend to bias towards larger families, those with more free time to answer long surveys, and, of course, families with phones. Moreover, HCMST adopted a weighting scheme that oversampled same-sex couples, as well as other less-common demographics. This further suggests caution about generalizing our results to the general population.

We would be interested in additional data columns documenting the location of each couple, the amount of time they spend together, area of employment, and other details that could predict key features of their relationship status. We would also be interested in global data for comparing US relationship trends with that of other countries.

If we had time, a fascinating experiment would be to collect relationship information from students at Yale to see where we deviate from the trends in the HCMST sample, and whether it is possible to predict how long couples will last, based on our data.

5 Code Appendix

1. EXPLORATORY PLOTS

A. Histogram of Relationship Longevity

```r
hcmst.cut <- hcmst
#remove NAs from relationship longevity variable:
rel_long <- hcmst.cut$HOW_LONG_RELATIONSHIP[which(!is.na(hcmst.cut$HOW_LONG_RELATIONSHIP))]
#length(rel_long) #2983 respondents
mean_rl <- mean(rel_long)
median_rl <- median(rel_long)
data.frame(year = rel_long) %>% ggplot(aes(x = year)) +
  geom_histogram(bins = 20, fill = 'white', color = 'black') +
  geom_segment(aes(x = mean_rl, y = 0, xend = mean_rl, yend = 510,
                   color = 'Mean (17.7 yrs)')) +
  geom_segment(aes(x = median_rl, y = 0, xend = median_rl, yend = 510,
                   color = 'Median (13 yrs)')) +
  ggtitle('Histogram of Relationship Longevity') +
  xlab('Relationship Longevity (in years)') + ylab('Counts') +
  scale_colour_discrete(name = "Color")

#hist(rel_long, breaks = 50, main = "Histogram of Relationship Longevity",
#     xlab = "Relationship Longevity (in years)", ylab = "Counts")
#abline(v = c(mean_rl, median_rl), col = c(2,3), lwd = 3)
#legend("topright", c("Mean (17.7 yrs)", "Median (13 yrs)"),
#       pch = "-", col = c(2,3), lwd = c(3,3))
quantile(rel_long, seq(0,1,0.1))
```

B. Parental Approval and Relationship Quality

```r
input <- hcmst.cut[!is.na(hcmst.cut$PARENTAL_APPROVAL) & !is.na(hcmst.cut$RELATIONSHIP_QUALITY),]
quality <- factor(input$RELATIONSHIP_QUALITY,
                  labels = c("Very Poor", "Poor","Fair","Good","Excellent"))
approval <- factor(input$PARENTAL_APPROVAL, labels = c("No/Unknown","Yes"))
ggplot(data = input, aes(approval,
  fill = quality)) +
  geom_bar(stat="count", position = "dodge") +
  xlab('Parental Approval') + ylab('Counts') +
  guides(fill=guide_legend(title='Relationship Quality')) +
  ggtitle('Relationship Quality Depends on Parental Approval')
```

C. Importance of Attractiveness by Gender

```r
input <- input[!is.na(input$W4_ATTRACTIVE_PARTNER) & !is.na(input$PPGENDER),]
ggplot(data = input, aes(factor(input$PPGENDER, labels = c('Male','Female')),
  fill = factor(input$W4_ATTRACTIVE_PARTNER))) +
  geom_bar(stat="count", position = "dodge") + xlab('Gender') + ylab('Counts') +
  guides(fill=guide_legend(title='Rated Attractiveness of Partner')) +
  ggtitle('Rated Partner Attractiveness by Gender')
```

2. ANALYSIS

A. Methods of Meeting and the Rise of Online Dating

```r
#1: heterosexual, 2: homosexual couple
```

```
hcmst$YEAR_MET <- as.integer(2009 - hcmst$HOW_LONG_AGO_FIRST_ROMANTIC)
hcmst.cutmeet<- hcmst[!is.na(hcmst$SAME_SEX_COUPLE),]
heterosexual <- hcmst.cutmeet[hcmst.cutmeet$SAME_SEX_COUPLE == 1,]
homosexual <- hcmst.cutmeet[hcmst.cutmeet$SAME_SEX_COUPLE == 2,]

homomeet <- data.frame(year = seq(1980, 2009, 1), online = rep(0,30),
                       cowork = rep(0,30), friend = rep(0,30),
                       family = rep(0,30), neighbor = rep(0,30),
                       college = rep(0,30), school = rep(0,30),
                       church = rep(0,30), bar = rep(0,30))

for (k in 28:36){
  for (i  in 1: length(homosexual[,1])){
    if (!is.na(homosexual$YEAR_MET[i])){
      if(homosexual[i,k] > 0 && !is.na(homosexual[i,k])){
        index <- which(homomeet$year==homosexual$YEAR_MET[i])
        homomeet[index,(k-26)] = homomeet[index,(k-26)] + 1
      }
    }
  }
}

heteromeet <- data.frame(year = seq(1980, 2009, 1), online = rep(0,30),
                       cowork = rep(0,30), friend = rep(0,30),
                       family = rep(0,30), neighbor = rep(0,30),
                       college = rep(0,30), school = rep(0,30),
                       church = rep(0,30), bar = rep(0,30))

for (k in 28:36){
  for (i  in 1: length(heterosexual[,1])){
    if (!is.na(heterosexual$YEAR_MET[i])){
      if(heterosexual[i,k] > 0 && !is.na(heterosexual[i,k])){
        index <- which(heteromeet$year==heterosexual$YEAR_MET[i])
        heteromeet[index,(k-26)] = heteromeet[index,(k-26)] + 1
      }
    }
  }
}

homomeet$yearsum = rowSums(homomeet[,2:10])
heteromeet$yearsum = rowSums(heteromeet[,2:10])

homomeet <- homomeet %>% mutate(online = online/yearsum) %>%
  mutate(cowork = cowork/yearsum) %>%
  mutate(friend = friend/yearsum) %>%
  mutate(family = family/yearsum) %>%
  mutate(neighbor = neighbor/yearsum) %>%
  mutate(college = college/yearsum) %>%
  mutate(school = school/yearsum) %>%
  mutate(church = church/yearsum) %>%
  mutate(bar = bar/yearsum) %>%
  select(-yearsum)
```

```
modeofmeeting <- c("online","cowork","friend","family",
                   "neighbor","college","school","church","bar")

# Apply loess function
homomeet[,2:10]<- sapply(2:10, function(x) loess(homomeet[,x] ~ homomeet$year)$fitted)
homomeetmelt <- melt(homomeet, id.vars="year", value.name="value")

heteromeet <- heteromeet %>% mutate(online = online/yearsum) %>%
  mutate(cowork = cowork/yearsum) %>%
  mutate(friend = friend/yearsum) %>%
  mutate(family = family/yearsum) %>%
  mutate(neighbor = neighbor/yearsum) %>%
  mutate(college = college/yearsum) %>%
  mutate(school = school/yearsum) %>%
  mutate(church = church/yearsum) %>%
  mutate(bar = bar/yearsum) %>%
  select(-yearsum)


# Apply loess function
heteromeet[,2:10]<- sapply(2:10, function(x)
  loess(heteromeet[,x] ~ heteromeet$year)$fitted)

#melt variables
heteromeetmelt <- melt(heteromeet, id.vars="year", value.name="value")

homomeetmelt$samesex <- 'Same-sex'
heteromeetmelt$samesex <- "Opposite-sex"
dfmeet <- rbind(homomeetmelt, heteromeetmelt)
ggplot(data=dfmeet, aes(x=year, y = value, group = variable, color = variable)) +
  geom_line() + geom_point(aes(shape=variable))+
  ggtitle ("Method of Meeting for HCMST Couples") +
  ylab("Fraction of Couples")+xlab("Year Met") +facet_wrap(~samesex) +
  theme(panel.spacing = unit(1.5, "lines")) +
  scale_color_discrete(name = 'Met') + scale_shape_discrete(name = 'Met')
```

B. The Demographics of Online Meeting

Politics

```
hcmst.online <- hcmst %>% filter(!is.na(MET_ONLINE))

table.prop=function(x) {
  tmp=table(x)
    tmp.a=apply(tmp,2,sum) #sum along the columns
    t(tmp)/tmp.a #transpose the table and divide by the sum to give percentage
}

output <- table.prop(hcmst.online[,c(10,28)])
rownames(output) <- c('Met Offline', 'Met Online')
output %>% signif(2) %>% kable
chisq.test(table(hcmst.online[,c(10,28)])) %>% invisible
```

Education

```
output <- table.prop(hcmst.online[,c(8,28)])
rownames(output) <- c('Met Offline', 'Met Online')
colnames(output) <- c('Less than high school','High school',
                      'Some college','Bachelor’s degree')
output %>% signif(2) %>% kable
chisq.test(table(hcmst.online[,c(8,28)])) %>% invisible
```

Parental Approval

```
output <- table.prop(hcmst.online[,c(21,28)])
rownames(output) <- c('Met Offline', 'Met Online')
colnames(output) <- c('Disapproval or Unknown','Approval')
output %>% signif(2) %>% kable
chisq.test(table(hcmst.online[,c(21,28)])) %>% invisible
```

C. The Role of Religion and Race in Partner Selection

Respondent Filtering

```
sum(is.na(hcmst$RESPONDENT_RELIG_16_CAT))/nrow(hcmst)
# Are the missing values just no partners?
nrow( hcmst %>% filter(is.na(PARTNER_RELIG_16_CAT)))
nrow( hcmst %>% filter(is.na(PARTNER_RACE)))
nrow( hcmst %>% filter(is.na(PARTNER_YRSED)))
nrow( hcmst %>% filter(is.na(PARTNER_RELIG_16_CAT) &
                       is.na(PARTNER_RACE) & is.na(PARTNER_YRSED)))
hcmst <- hcmst %>% filter(!is.na(PARTNER_RELIG_16_CAT) &
                          !is.na(PARTNER_RACE) & !is.na(PARTNER_YRSED))
nrow(hcmst)
hcmst <- hcmst %>% filter(!is.na(PARTNER_RELIG_16_CAT) &
                          !is.na(PARTNER_RACE) & !is.na(PARTNER_YRSED))
```

D. The Role of Religion in Partner Selection

```
hcmst$RELIGION_SAME <- 0
hcmst$RELIGION_SAME[which(hcmst$PARTNER_RELIG_16_CAT==hcmst$RESPONDENT_RELIG_16_CAT)] <- 1

x <- sum(hcmst$RELIGION_SAME)/nrow(hcmst)

#double checking that the percentage of people with the same religion
#is the same using the brute force method (here) as the above method
#using the RELIGION_SAME variable
table(hcmst$RELIGION_SAME, exclude=NULL)
table(hcmst$PARTNER_RELIG_16_CAT,hcmst$RESPONDENT_RELIG_16_CAT)
(table(hcmst$PARTNER_RELIG_16_CAT,hcmst$RESPONDENT_RELIG_16_CAT)[1,1]+
    table(hcmst$PARTNER_RELIG_16_CAT,hcmst$RESPONDENT_RELIG_16_CAT)[2,2]+
    table(hcmst$PARTNER_RELIG_16_CAT,hcmst$RESPONDENT_RELIG_16_CAT)[3,3]+
    table(hcmst$PARTNER_RELIG_16_CAT,hcmst$RESPONDENT_RELIG_16_CAT)[4,4]+
    table(hcmst$PARTNER_RELIG_16_CAT,hcmst$RESPONDENT_RELIG_16_CAT)[5,5])/nrow(hcmst)

n <- 10000
frac_same_holder <- rep(NA, n)
temp.hcmst <- data.frame(RESPONDENT_RELIG_16_CAT = rep(NA, nrow(hcmst)),
                         PARTNER_RELIG_16_CAT = rep(NA, nrow(hcmst)) ,
                         RELIGION_SAME = rep(NA, nrow(hcmst)),
                         RELIGION_SAME_TYPE = rep(NA, nrow(hcmst)))
```

```
temp.hcmst$RESPONDENT_RELIG_16_CAT <- hcmst$RESPONDENT_RELIG_16_CAT
set.seed(230)

# Actual simulation
for (i in 1:n) {
   # 1. Create sample
   temp.hcmst$PARTNER_RELIG_16_CAT <- sample(hcmst$PARTNER_RELIG_16_CAT, replace = FALSE)
   # 2. Generate derived columns.
   temp.hcmst$RELIGION_SAME <- 0
   assign <- which(temp.hcmst$PARTNER_RELIG_16_CAT==temp.hcmst$RESPONDENT_RELIG_16_CAT)
   temp.hcmst$RELIGION_SAME[assign] <- 1
   frac_same_holder[i] <- sum(temp.hcmst$RELIGION_SAME)/nrow(temp.hcmst)
}
mean(frac_same_holder)
# Relative difference
(sum(hcmst$RELIGION_SAME)/nrow(hcmst))/mean(frac_same_holder)-1
c <- quantile(frac_same_holder, c(0.025,0.975))
c %>% t %>% kable
hist(frac_same_holder, breaks = 50, xlim = c(0.35,0.6), cex = 1,
     main = "Matching Religions Between Respondents and Permuted Partners",
     xlab="Fraction of couples with same religion", ylab = "Counts")
abline(v=c, lwd=2,col="blue")
abline(v=x, lwd=2,col="red")
legend("topright", c("95% confidence interval",
       "observed fraction of couple religion matches"),
       lwd = c(2,2), col = c("blue", "red"))
```

E. Online vs. Offline Selection of Same Religious Partners

```
hcmst.online.2 <- hcmst %>% filter(!is.na(MET_ONLINE))
output <- table.prop(hcmst.online.2[,c(40,28)])
rownames(output) <- c('Met Offline', 'Met Online')
colnames(output) <- c('Different Religion','Same Religion')
output %>% signif(2) %>% kable
chisq.test(table(hcmst.online.2[,c(40,28)])) %>% invisible
```

F. The Role of Race in Partner Selection

```
# Binary for Same Race
hcmst$RACE_SAME <- 0
hcmst$RACE_SAME[which(hcmst$RESPONDENT_RACE==hcmst$PARTNER_RACE)] <- 1
x <- sum(hcmst$RACE_SAME)/nrow(hcmst)

table(hcmst$PARTNER_RACE,hcmst$RESPONDENT_RACE)
(table(hcmst$PARTNER_RACE,hcmst$RESPONDENT_RACE)[1,1]+
    table(hcmst$PARTNER_RACE,hcmst$RESPONDENT_RACE)[2,2]+
    table(hcmst$PARTNER_RACE,hcmst$RESPONDENT_RACE)[3,3]+
    table(hcmst$PARTNER_RACE,hcmst$RESPONDENT_RACE)[4,4]+
    table(hcmst$PARTNER_RACE,hcmst$RESPONDENT_RACE)[5,5]+
    table(hcmst$PARTNER_RACE,hcmst$RESPONDENT_RACE)[6,6])/nrow(hcmst)

# Boot permutation test.
# Preparation
n <- 10000
frac_same_race_holder <- rep(NA, n)
```

```
temp.hcmst <- data.frame(RESPONDENT_RACE = rep(NA, nrow(hcmst)),
                         PARTNER_RACE = rep(NA, nrow(hcmst)) ,
                         RACE_SAME = rep(NA, nrow(hcmst)),
                         RACE_SAME_TYPE = rep(NA, nrow(hcmst)))
temp.hcmst$RESPONDENT_RACE <- hcmst$RESPONDENT_RACE
set.seed(230)

# Simulation
for (i in 1:n) {
   # 1. Create sample
   temp.hcmst$PARTNER_RACE <- sample(hcmst$PARTNER_RACE, replace = FALSE)
   # 2. Generate derived columns.
   temp.hcmst$RACE_SAME <- 0
   assign <- which(temp.hcmst$PARTNER_RACE==temp.hcmst$RESPONDENT_RACE)
   temp.hcmst$RACE_SAME[assign] <- 1
   frac_same_race_holder[i] <- sum(temp.hcmst$RACE_SAME)/nrow(temp.hcmst)
}

# Observed fraction with same religion
x <- sum(hcmst$RACE_SAME)/nrow(hcmst)
x

mean(frac_same_race_holder)

# Relative difference
(sum(hcmst$RACE_SAME)/nrow(hcmst))/mean(frac_same_race_holder)-1

# Finding confidence interval
c <- quantile(frac_same_race_holder, c(0.025,0.975))
c %>% t %>% kable

# Histogram

hist(frac_same_race_holder, breaks = 50, xlim = c(0.6,0.85), cex = 1,
     main = "Matching Race Between Respondents and Permuted Partners",
     xlab="Fraction of couples with same race", ylab = "Counts")
abline(v=c, lwd=2,col="blue")
abline(v=x, lwd=2,col="red")
legend("topright", c("95% confidence interval",
       "observed fraction of couple race matches"),
       lwd = c(2,2), col = c("blue", "red"))

unique(hcmst$RESPONDENT_RACE)

#white couples within race
table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[1,1]/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[1,1:6])
# Outside
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[1,2:6])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[1,1:6])
# Whites as fraction of population
nrow(hcmst[which(hcmst$RESPONDENT_RACE=="(1) NH white"),])/nrow(hcmst)
```

```
#Black Within
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[2,2])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[2,1:6])
#Outside
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[2,c(1,3:6)])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[2,1:6])
# Fraction of Blacks
nrow(hcmst[which(hcmst$RESPONDENT_RACE=="(2) NH black"),])/nrow(hcmst)

#American Indian
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[3,3])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[3,1:6])
#Outside
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[3,c(1:2,4:6)])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[3,1:6])
#Fraction
nrow(hcmst[which(hcmst$RESPONDENT_RACE=="(3) NH Amer Indian"),])/nrow(hcmst)

#Asian
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[4,4])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[4,1:6])
#Outside
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[4,c(1:3,5:6)])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[4,1:6])
# Fraction
nrow(hcmst[which(hcmst$RESPONDENT_RACE=="(4) NH Asian Pac Islander"),])/nrow(hcmst)

#Hispanic
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE,exclude=NULL)[6,6])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[6,1:6])
#Outside
sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[6,c(1:5)])/
  sum(table(hcmst$RESPONDENT_RACE,hcmst$PARTNER_RACE, exclude=NULL)[6,1:6])
# Fraction of sample
nrow(hcmst[which(hcmst$RESPONDENT_RACE=="(6) Hispanic"),])/nrow(hcmst)
```