SML 201 Project 1

Tyler Campbell

March 21, 2018

Project 1 is due by 11:59p.m. on Saturday March 17 on Blackboard. Please submit both a .Rmd and a .pdf file on Blackboard. Please drop off the pdf copy at 26 Prospect Avenue (see the *Submitting Problem Sets and Projects* section under *Problem Sets and Projects* on the Syllabus for detailed instructions) as soon as you can and no later than Monday March 26 5 p.m. Please do not modify your .pdf hard copy after your BB submission; otherwise, you might get points deducted.

Late **projects** will be penalized at intervals rounded up to multiples of 24 hours. For example, if you are 3 hours late, 10% off or if you are 30 hours late, 20% off.

Make sure that you have all your digital signatures along with the honor pledge in each of these documents (there should be more than one signature if you work in groups).

This project can be completed in groups of up to 3 students. It is okay to work by yourself, if this is preferable. **You may not work with a given student on more than one project.** In other words, if you work with Student_1 and Student_2 on Project 1, then you cannot work with Student_1 or Student_2 on any other projects. You must form completely new groups for every project.

When working in a group it is your responsibility to make sure that you are satisfied with all parts of the report and the submission is on time (e.g., we will not entertain arguments that deficiencies are the responsibility of other group members). We expect that the work on any given problem set or project contains approximately equal contributions from all members of the group; we expect that you each work independently first and then compare your answers with each other once you all finish or you all work together. Failing to make contributions and then putting your name on a project will be considered a violation of the honor code. Also, please do not divide work among your group mates.

In general you are not allowed to get help on projects from other people except from partners in your group; there is an exception for this project: you are allowed to get help from instructors if you need help to understand the definitions of the variables for the dataset or the procedure of the experiment. Clarification questions are always welcome. Please treat projects as take-home exams.

For all parts of this problem set, you **MUST** use R commands to print the output as part of your R Markdown file. You are not permitted to find the answer in the R console and then copy paste the answer into this document.

If you are completing this problem set in a group, please have only **one** person in your group turn in the .Rmd and .pdf files; other people in your group should turn in the list of the people in your group in the *Text Submission* field on the submission page.

Please type your name(s) after "Digitally signed:" below the honor pledge to serve as digital signature(s). Put the pledge and your signature(s) at the beginning of each document that you turn in.

> I pledge my honor that I have not violated the honor code when completing this assignment.

Digitally signed: Tyler Campbell

In order to receive full credits, please have sensible titles and axis labels for all your graphs and adjust values for all the relevant graphical parameters so that your plots are informative. Also, all answers must be written in complete sentences.

Just a friendly reminder: Please remember to annotate your code and have answers in the write up section, not in code chunks.

Objective of this project

In this project we will explore a subset of the data that were gathered on participants in 21 experimental speed dating events that took place between 2002 and 2004. The goal of this project is to investigate whether there is a gender difference in terms of behavior of participants.

Background info on this project

The dataset for this project is a subset of the data published in the paper "Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment" by Fisman et al (due to data quality issues our data have been modified slightly).

How were the experiments designed?

We have included a copy of the paper where the original dataset was published in (see *Gender_Differences_in_Mate_Selection.pdf*). Please read the part of the *EXPERIMENTAL DESIGN AND DATA DESCRIPTION* section that is on page 676-678 and answer the following question.

Question 1

Who are the participants in the study? Note that this is a very specific population so it might not be appropriate to generalize the results that we find in this project to a different population.

Participants were students in graduate and professional schools at Columbia University between 2002-2004. (All students willfully registered)

Getting familiar with the dataset

Question 2

Part a

The file `Speed Dating Data Key.doc` includes the variable definitions for variables in the original dataset for the paper. Please read through this document; it should familiarize you with the design of the experiment. List the definitions/meanings of the variables in the dataset for this project. No code required for this part. iid: unique subject number, group(wave id gender)

gender: Female=0, Male=1

wave: wave number

match: 1=yes, 0=no

dec_o: decision of partner the night of event

age: age of the participant

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?

When looking at what the partcipate looks for in 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

dec: decsision to indicate whether or not the particapant would like to other person again

Rate their attributes on a scale of 1-10: (1=awful, 10=great), N/A if they can't come up with score

attr: Attractive, sinc: Sincere, intel: Intelligent, fun: Fun, amb: Ambitious, shar: Has shared interests/hobbies,

prob: How probable do you think it is that the other person will say 'yes' for the participant? (1=not probable, 10=extremely probable)

Part b

Read in the data from `Speed Dating Subset.csv` and name this object as `dating`. What is the object type of `dating` and what are the dimensions of `dating`? objectype: dataframe

dimensions: 8378 rows by 21 columns

```
# read in file into dataframe, check dimensions and
# object type
dating = read.csv("Speed Dating Subset.csv")
class(dating)
[1] "data.frame"
dim(dating)
[1] 8378   21
```

Part c

The data of only 14 speed dating sessions were used (see table I on page 678) in the paper. How many sessions does our dataset `dating` cover?

21

```
# finds the amount of unique wav numbers to find
# the number of sessions
length(unique(dating$wave))
[1] 21
```

Demographics of the participants

Problem 3

Explore the dataset `dating` to answer the questions in this problem.

Part a

How many (distinct) participants were recorded in this dataset? How many of them are male and how many of them are female?

551 participants. 277 Males and 274 Females

```
# find the length of unique iid values to find
# number of participants
length(unique(dating$iid))
[1] 551
# remove duplicate iid numbers from dataframe to
# isolate 1 participant
dt = dating[!duplicated(dating$iid), ]
# create table from number of male and females
```

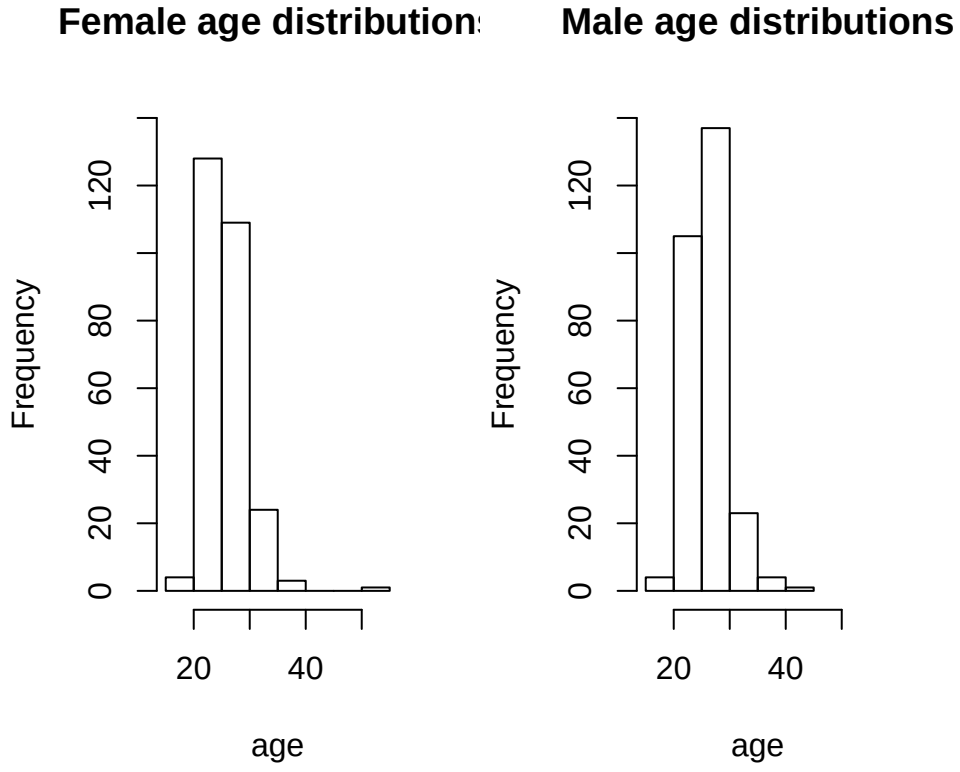
```
table(dt$gender)

  0   1
274 277
```

Part b

Make two histograms, one for the age distribution of each gender, to compare the age distributions of the male and female participants. You will only be reminded of this once for the entire course so please remember to do this in the future too: please make sure that the histograms are on the same axis scales and ranges for the comparison.

```
# creates histograms of the age distributions of
# the male and female participants
f = dt[dt$gender == 0, ]
m = dt[dt$gender == 1, ]
par(mfrow = c(1, 2))
hist(f$age, xlim = c(15, 55), ylim = c(0, 140), xlab = "age",
    main = "Female age distributions")
hist(m$age, xlim = c(15, 55), ylim = c(0, 140), xlab = "age",
    breaks = 5, main = "Male age distributions")
```



Do male and female behave differently in these events?

Problem 4

Part a

According to the survey instructions (see file `Speed Dating Data Key.doc`) what are the possible values the participants are supposed to give for the variable `prob`? Check your dataset `dating` and what are the possible values for `prob` in the dataset? It is not uncommon that participants failed to follow the instructions for a survey. We will fix this by rounding the values for `prob` to integers with the `round()` function; there will still be zero values after the rounding and we will just keep them. Name the vector with the rounded values of `prob` as `r.prob` and use `r.prob` instead of `prob` for the rest of the problem.

Participants are supposed to give values integer values 1-10. Possible values 6.0 5.0 NA 7.0 4.0 3.0 8.0 1.0 10.0 2.0 9.0 4.5 6.5 5.5 0.0 7.5 8.5 9.5 3.5 1.5. Out of range and non integer values.

```
# make rounded vector of prob and add it to the
# dating dataframe
unique(dating$prob)
 [1]  6.0  5.0   NA  7.0  4.0  3.0  8.0  1.0 10.0  2.0  9.0  4.5  6.5  5.5
[15]  0.0  7.5  8.5  9.5  3.5  1.5
r.prob = round(dating$prob)
dating$r.prob = r.prob
```

Part b

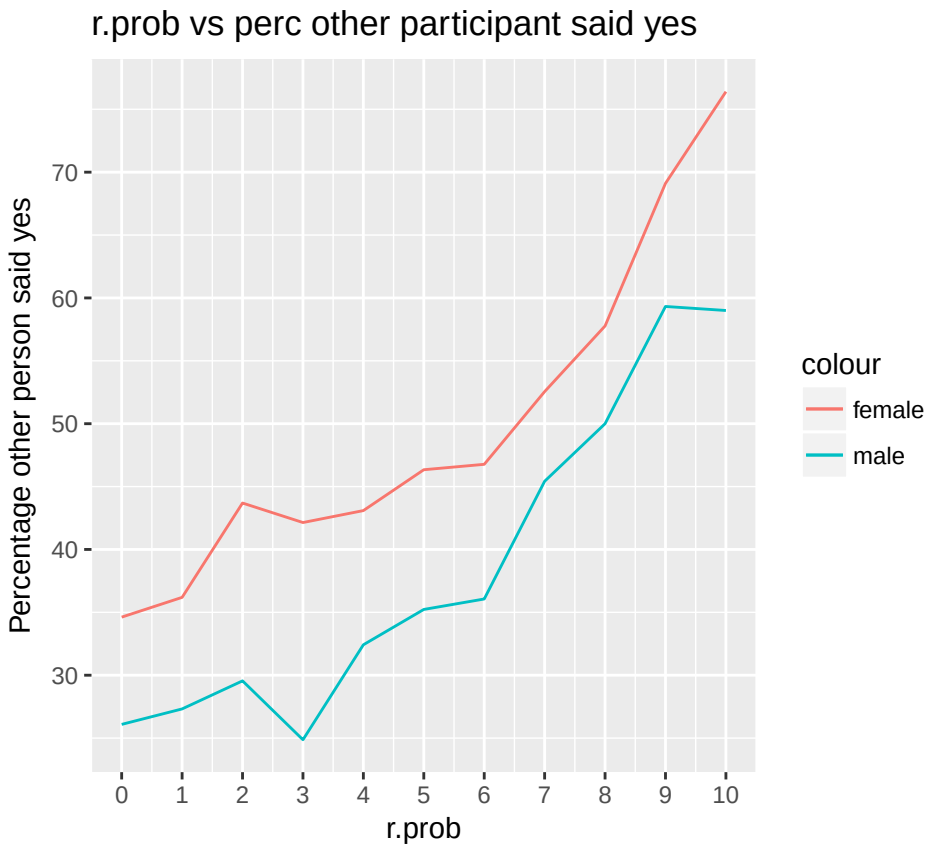
Make a line plot; the plot should have two lines, one for each gender. The x-coordinate of the plot should be for the values of `r.prob` and the y-coordinate should be for the percentage of the times the partner of a participant (who gave the partner a rating x for `r.prob`) would like to see the participant again. Compare the two lines and answer this: Say, female participant Kate and male participant Kent both give a rating x (x takes on the value in {0, 1, 2, . . . , 10}) for how probable their partners will say "yes" to them; according to the graph that you just made, is it more likely for Kate's or Kent's partner to say "yes" to her/him? Is this result consistent for all possible values of x?

Kate's partner is more probalble to say yes. Yes, this result is consistent for all possible values of x.

```
# male lineplot with each point reprsenting the
# prob a participants gives and the percentage the
# other person said ues to meet again. Two lines
# represnt female and male
library(ggplot2)
female = dating[dating$gender == 0, ]
male = dating[dating$gender == 1, ]
# calculate perctanges for each r.prob
```

```r
i = seq(2, 22, 2)
fyes = table(female$dec_o, female$r.prob)[i]
ftotal = as.vector(table(female$r.prob))
myes = table(male$dec_o, male$r.prob)[i]
mtotal = as.vector(table(male$r.prob))
rate = sort(unique(r.prob))
fperc = round(fyes/ftotal * 100, digits = 2)
mperc = round(myes/mtotal * 100, digits = 2)

df = data.frame(rate, fperc, mperc)
# line plot
ggplot(df, aes(x = rate)) + geom_line(aes(y = fperc,
    colour = "female")) + geom_line(aes(y = mperc,
    colour = "male")) + scale_x_continuous(breaks = 0:10) +
    labs(x = "r.prob", y = "Percentage other person said yes",
        title = "r.prob vs perc other participant said yes")
```

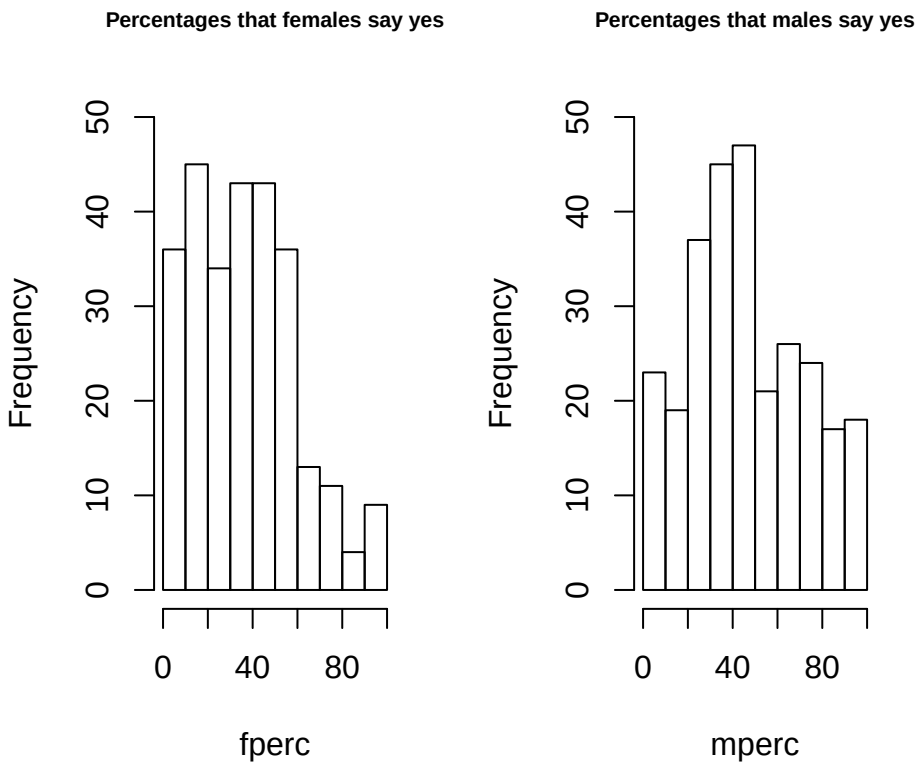


Part c

Are men or women more likely to say "yes" (i.e., `dec = 1`) to their speed dating partners? Make two histograms, one for each gender; the histograms should display the distribution of the percentage of times a participant said "yes" that he/she would like to see his/her speed dating partner again. You should have one data point for each participant; for example, for participant with `iid` 1 you should find out the percentage of the times participant 1 said "yes" to her speed dating partners.

Men are more likely to say yes.

```
# calculates the percantages yes for each female
# participant
fiid = as.vector(table(female$iid))
fyes = table(female$iid, female$dec)[275:548]
fperc = round(fyes/fiid * 100, digits = 2)
# calculates the percantages yes for each male
# participant
miid = as.vector(table(male$iid))
myes = table(male$iid, male$dec)[278:554]
mperc = round(myes/miid * 100, digits = 2)
# creates histogram of perctages yes for both
# genders
par(mfrow = c(1, 2))
hist(fperc, ylim = c(0, 50), main = "Percentages that females say yes",
    cex.main = 0.7)
hist(mperc, ylim = c(0, 50), main = "Percentages that males say yes",
    cex.main = 0.7)
```



Part d

Report the 5 number summary for each of the distributions in part c. Answer the question proposed in part c: Among the participants are men or women more likely to say "yes" to their speed dating partners? Among the male participants what is the percentage of men who could not find any dates that they would like to see again, and what is the percentage of men who would like to see every single one of their dates that they met again? Answer these two questions for the female participants.

Min. 1st Qu. Median 3rd Qu. Max.

for female: 0.00 18.75 36.84 54.24 100.00

for male: 0.00 30.00 44.44 66.67 100.00

Men are more likey to say yes.

3.25 % of males could not find any dates that they would like to see again.

5.05 % of males would like to see every single one of their dates that they met again

8.03 % of females could not find any dates that they would like to see again.

2.19% of females would like to see every single one of their dates that they met again

```
# fivenumber summary for each male and female
# percentages
```

```
summary(fperc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   18.75   36.84   37.38   54.24  100.00
summary(mperc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   30.00   44.44   48.11   66.67  100.00
# calculates the percantage of male and females
# that said all no or all yes
numMperc = as.vector(table(mperc))
numFperc = as.vector(table(fperc))
allnoM = numMperc[1]
allyesM = numMperc[length(numMperc)]
allnoF = numFperc[1]
allyesF = numFperc[length(numFperc)]
round(allnoM/length(mperc) * 100, digits = 2)
[1] 3.25
round(allyesM/length(mperc) * 100, digits = 2)
[1] 5.05
round(allnoF/length(fperc) * 100, digits = 2)
[1] 8.03
round(allyesF/length(fperc) * 100, digits = 2)
[1] 2.19
```

Question 5

Part a

For this question we will use the observations without missing data only. Subset the variables "iid", "gender", "match", "dec_o", "exphappy" and "dec" from the dataset `dating` and call this subset `sub`. Use `na.omit()` to extract out the observations in `sub` with complete data only. Name the subset with complete observations as `compl.date`. What percentage of the observations are removed by `na.omit()`? Note that we are removing only a small fraction of our observations. Use `compl.date` for the rest of the problem.

1.21 % of observations are removed

```
# create compl.date by removing na's and subset
# from correct varibles
sub = subset(dating, select = c("iid", "gender", "match",
    "dec_o", "exphappy", "dec"))
compl.date = na.omit(sub)
# calculate percent removed
a = 1 - (dim(compl.date)[1]/dim(sub)[1])
round(a * 100, digits = 2)
[1] 1.21
```

Part b

Were the values for the variable `exphappy` recorded before, during, or after a speed dating event? Note that the value for `exphappy` was only recorded once in time so this value should be identical for all entries for a given participant.

exphappy is before the speed dating event

Part c

We would like to check the quality of the data and we want to see if the values for `exphappy` are identical for all the data entries for a particular participant. Given a participant we can calculate a numerical summary (which should be a single number) of the `exphappy` values associated with this participant to check if all the `exphappy` values are identical for this participant. What is this numerical summary? What value do you expect to see for this numerical summary if all the `exphappy` values are identical for a participant? Calculate this numerical summary for all participants to confirm that the `exphappy` values associated with each participant are the same for each participant. Make a vector `ind.exphappy` to store the `exphappy` value for each participant; the length of `ind.exphappy` should be 543 since there should be one entry for each participant.

variance, if all numbers are the same the value should be 0

```
# get all unique ids which represents each
# partipicant
participants = unique(compl.date$iid)
ind.exphappy = c()
# for every partipicant check that exphappy is
# consistent and add value to vector ind.exphappy
for (x in participants) {
    a = compl.date[compl.date$iid == x, ]$exphappy
    if (var(a) != 0)
        print("mistake")
    ind.exphappy = append(ind.exphappy, mean(a))
}
```

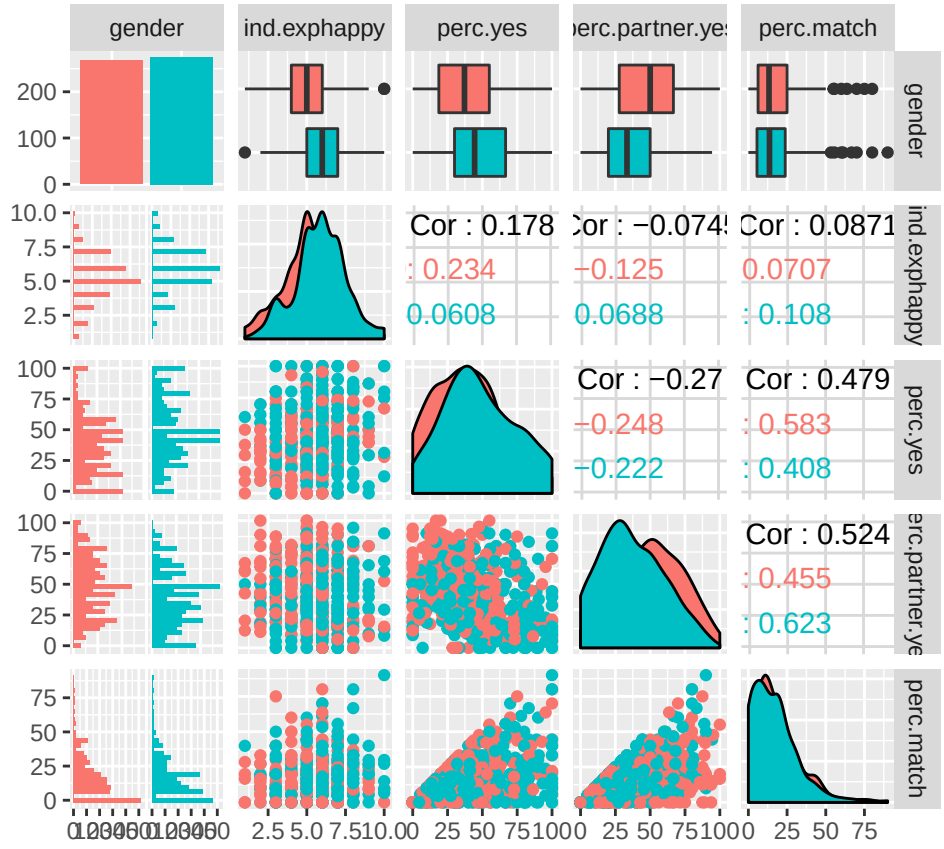
Part d

Create the following variables and use `ggpairs()` in the `GGally` package to make a matrix of plots to investigate the pairwise relationship between the variables *within each gender group*:

- gender: The gender of the participant;
- ind.exphappy: The `exphappy` value for each speed dating participant;
- perc.yes: The percentage of the dates the participant wanted to see again;
- perc.partner.yes: The percentage of the dates who would like to see the participant again;
- perc.match: The percentage of matches the participant had.

```
library(GGally)
sex = c()
perc.yes = c()
perc.partner.yes = c()
perc.match = c()
# for every partipicant calculate
# gender,perc.yes,perc.partner.yes perc.match and
# add to a vector
for (x in participants) {
    p = compl.date[compl.date$iid == x, ]
    length = dim(p)[1]
    yes = sum(p$dec)
    partyes = sum(p$dec_o)
    match = sum(p$match)

    sex = append(sex, p$gender[1])
    perc.yes = append(perc.yes, round(yes/length *
        100, digits = 2))
    perc.partner.yes = append(perc.partner.yes, round(partyes/length *
        100, digits = 2))
    perc.match = append(perc.match, round(match/length *
        100, digits = 2))
}
gender = factor(sex)
data = data.frame(gender, ind.exphappy, perc.yes, perc.partner.yes,
    perc.match)
# ggpair for each varible
ggpairs(data, aes(colour = gender))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Part e

Based on your matrix of plots in part d answer following questions (no code required):

(i) The correlation between `ind.exphappy` and `perc.yes` is stronger for the group of female or male participants?

Stronger for females

(ii) A strong positive (or negative) linear relationship between the variables `exphappy` and `perc.match` would mean that people who felt optimistic (i.e., with high `exphappy` value) about their future potential speed dates are more (or less) likely to have matches. Based on your matrix of plots is there any strong linear pattern between the two variables `exphappy` and `perc.match`?

No strong linear pattern between exphappy and perc.match

Are people's professed thoughts consistent with their actions?

Problem 6

We will talk about this problem in Tuesday's lecture (discussion will not be covered in the notes so please make sure that you attend the lecture).

On the survey filled out before the speed dating events participants were asked to rank a set of qualities that they were looking for in the opposite sex. These qualities are: Attractiveness, sincerity, intelligence, fun, ambition, and having shared interests. In this problem we are trying to find out if people's professed thoughts are consistent with their selection criteria. Note that due to the large number of missing data for the variables that we use for this problem the result that we find might not be very reliable; nevertheless, we just want to have some fun and see what information we can extract out from the data.

Part a

Extract out the columns "iid", "gender", "attr1_1", "sinc1_1", "intel1_1", "fun1_1", "amb1_1", "shar1_1", "dec",
"attr", "sinc", "intel", "fun", "amb", "shar" from the dataset `dating` and divide the data into two sets: dataset `m` for the male participants and dataset `f` for the female participants.

```
# extract the given varibles and create 2
# dataframes for male and female
temp = subset(dating, select = c("iid", "gender", "attr1_1",
    "sinc1_1", "intel1_1", "fun1_1", "amb1_1", "shar1_1",
    "dec", "attr", "sinc", "intel", "fun", "amb", "shar"))
m = temp[temp$gender == 1, ]
f = temp[temp$gender == 0, ]
```

Part b

Extract out the variables "iid", "gender", "attr1_1", "sinc1_1", "intel1_1", "fun1_1", "amb1_1" and "shar1_1" from the dataset `m` and call this subset `male1`. "attr1_1", "sinc1_1", "intel1_1", "fun1_1", "amb1_1" and "shar1_1" are the variables that indicate how important these qualities are to a participant in terms what the participant was looking for in opposite sex (please see details in *Speed Dating Data Key.doc.*)

Make a 277 by 6 matrix `m.professed`. The columns of `m.professed` should correspond to the six personality quality variables mentioned in the previous paragraph. The rows of `m.professed` should give the values of the six personality quality variables for each male participant in `m`. (You can assume that there is no data quality issue; i.e., the values for a particular personality quality variable are always the same for multiple records that associate with the same participant.)

```r
# extract the wanted varibales and create matrix
# where a row represents a male particpiant and
# each column is one of the six personality quality
# variables
male1 = subset(m, select = c("iid", "gender", "attr1_1",
    "sinc1_1", "intel1_1", "fun1_1", "amb1_1", "shar1_1"))

money = male1[!duplicated(male1$iid), ]
money$iid = NULL
money$gender = NULL
m.professed = as.matrix(money)
```

Part c

Extract out the variables "iid", "gender", "dec",
"attr", "sinc", "intel", "fun", "amb" and "shar" from `m`; call this subset `male2`. Recall that `dec=1` means the participant would like to see his date again; you should also remind yourself the definitions of other variables in `male2`. We want to find out the average rating different between the rejected and the accepted groups of dates for each participant.

Create a 277 by 6 matrix `m.action`. The columns of `m.action` correspond to the six personality quality variables mentioned in the previous paragraph. The rows of `m.action` correspond to the 277 male participants. The entries correspond to the average rating different between the rejected group and the accepted group of dates of a participant; e.g., the (1,1) entry of `m.action` is

> average attr rating among the dates accepted by participant - average attr rating among the dates rejected by participant

for participant with `iid` 11.

```r
# extract the wanted varibales and create matrix
# where a row represents a male particpiant and
# each column is one of the six personality quality
# variables and each entry is average rating
# different between the rejected group and the
# accepted group of dates of a participant
male2 = subset(m, select = c("iid", "gender", "dec",
    "attr", "sinc", "intel", "fun", "amb", "shar"))
m.action = matrix(nrow = 277, ncol = 6)
i = 1
for (x in unique(male2$iid)) {
    p = male2[male2$iid == x, ]
    pyes = p[p$dec == 1, ]
    pno = p[p$dec == 0, ]
    for (y in 1:6) {
        z = y + 3
        m.action[i, y] = mean(pyes[, z]) - mean(pno[,
            z])
```

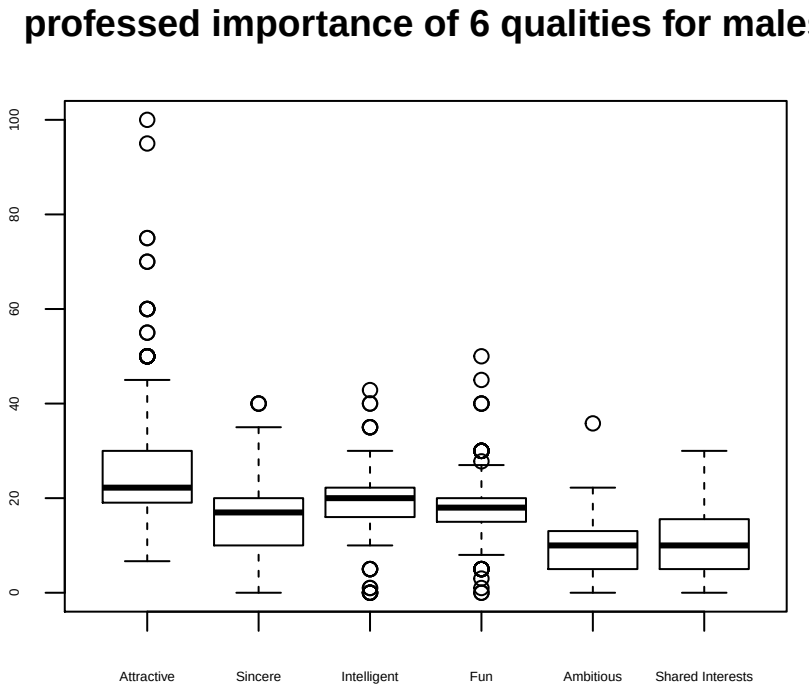
```
    }
    i = i + 1
}
```

Part d

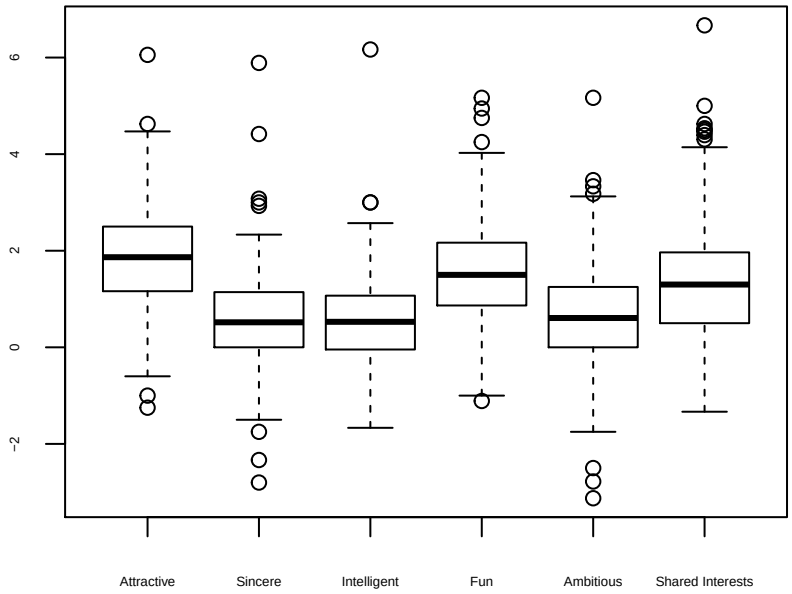
Make two side-by-side boxplots, one for each matrix that you created in parts b and c. Use 1st and 3rd quartiles as measures to answer the following questions.

```
# creates boxplots for professed and action of
# personality quality
colnames(m.professed) = c("Attractive", "Sincere",
    "Intelligent", "Fun", "Ambitious", "Shared Interests")
colnames(m.action) = c("Attractive", "Sincere", "Intelligent",
    "Fun", "Ambitious", "Shared Interests")

boxplot(m.professed, main = "professed importance of 6 qualities for males",
    cex.axis = 0.4)
```




```
boxplot(m.action, main = "in action importance of 6 qualities for males",
    cex.axis = 0.4)
```



Part e

Use 1st 2nd and 3rd quartiles as measures to answer the following questions.
Based on your plots in part d what are the top three qualities of a date the male participants think were the most important to them? What are the top three qualities that separate the rejected group from the accepted group of dates for the male participants? (No code required for this part.)

male participants think being attractive, fun and intelligent are most important

for male participants being attractive, fun and having shared interest are the top three qualities that separate the rejected group from the accepted group of dates

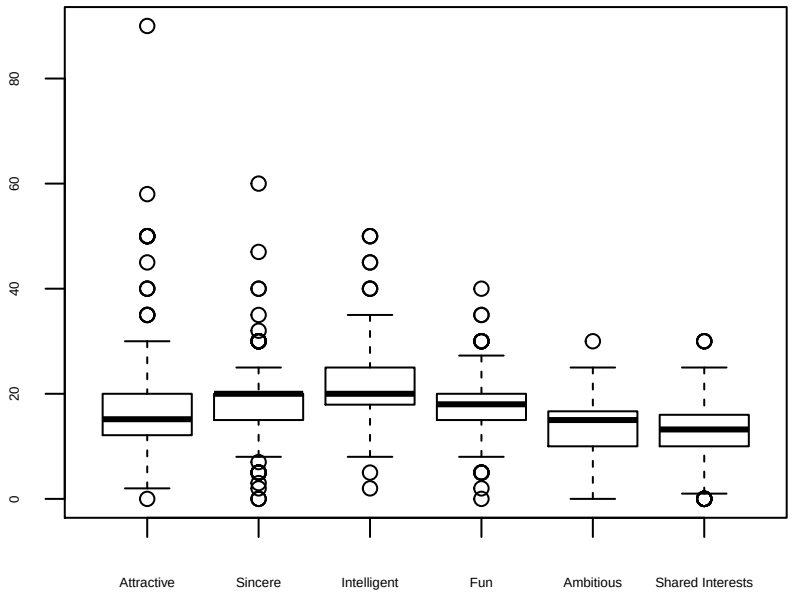
Part f

Modify your code and repeat parts b-d for the female participants.

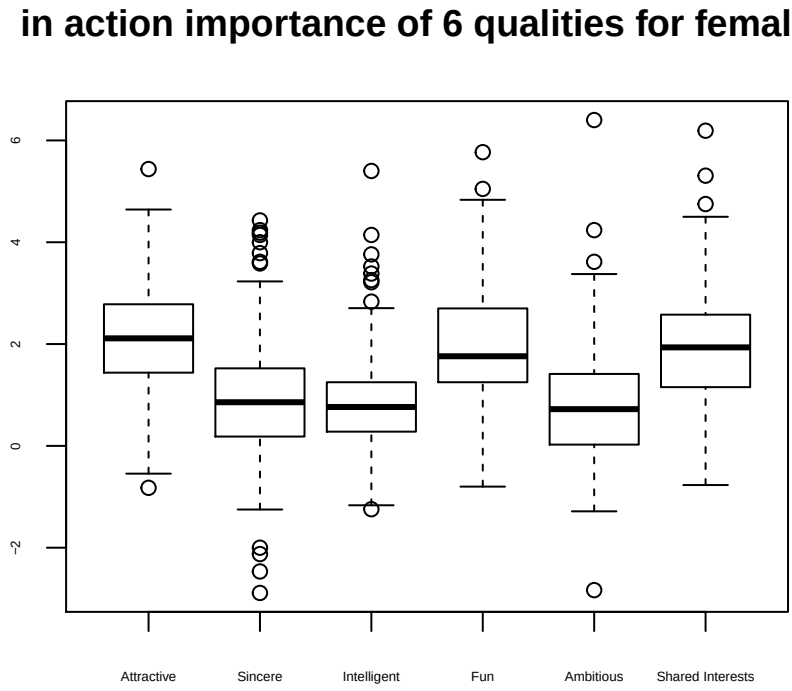
```
# creates same matrix in method ad before, but for
# females instead of males
female1 = subset(f, select = c("iid", "gender", "attr1_1",
    "sinc1_1", "intel1_1", "fun1_1", "amb1_1", "shar1_1"))

money = female1[!duplicated(female1$iid), ]
money$iid = NULL
money$gender = NULL
m.professed = as.matrix(money)
```

```r
female2 = subset(f, select = c("iid", "gender", "dec",
    "attr", "sinc", "intel", "fun", "amb", "shar"))
m.action = matrix(nrow = 274, ncol = 6)
i = 1
for (x in unique(female2$iid)) {
    p = female2[female2$iid == x, ]
    pyes = p[p$dec == 1, ]
    pno = p[p$dec == 0, ]
    for (y in 1:6) {
        z = y + 3
        m.action[i, y] = mean(pyes[, z]) - mean(pno[,
            z])
    }
    i = i + 1
}
# creates boxplots for professed and action of
# personality quality
colnames(m.professed) = c("Attractive", "Sincere",
    "Intelligent", "Fun", "Ambitious", "Shared Interests")
colnames(m.action) = c("Attractive", "Sincere", "Intelligent",
    "Fun", "Ambitious", "Shared Interests")
boxplot(m.professed, main = "professed importance of 6 qualities for females",
    cex.axis = 0.4)
```



```
boxplot(m.action, main = "in action importance of 6 qualities for females",
    cex.axis = 0.4)
```



Part g

Use 1st 2nd and 3rd quartiles as measures to answer the following questions.
Based on your plots in part f what is the top quality of a date the female participants think was the most important to them? What are the top three qualities that separate the rejected group from the accepted group of dates for the female participants? (No code required for this part.)

female participants think being sincere, fun and intelligent are most important

for female participants being attractive, fun and having shared interest are the top three qualities that separate the rejected group from the accepted group of dates