Report of Final Project

—Influence of 6 attributes in speed dating decision making

Yuzhi Sheng

Background

With the development of the society rapidly, the ways to make friends is changing, in another word, the demands of finding a romantic dating partner is increasingly growing. Therefore, as an efficient and fast way, a lot of single people are willing to take apart in speed dating.

Introduction

In this project, I will use a dataset which is from Kaggle.com named Speed Dating and select 6 attributes which are attractive, fun, sincere, intelligence, ambitious, share interest to analysis relationship between dating decision and these 6 attributes. Moreover, I will use some bar plots to calculate dating events, ages, races, study fields, dating goals and match or not. Also, I will show relationship between gender and study fields, gender and age and order and match. In addition, I would like to represent relationship between points and 6 attributes by using boxplot.

Data

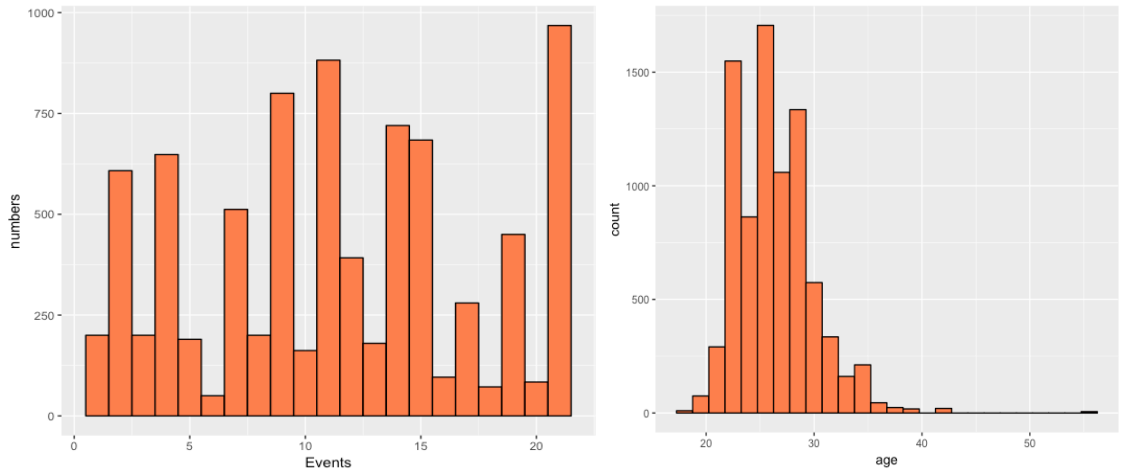
This dataset collects a lot of variables, but in this project, I choose several of them:
iid (unique subject number, group (wave id gender); gender (Female=0,Male=1); order (the number of date that night when met partner); partner (partner's id number the night of event); pid (partner's iid number); match (1=yes, 0=no); age_o (age of partner); race_o (race of partner); race: (1=Black/African American, 2=European/Caucasian-American, 3=Latino/Hispanic American, 4=Asian/Pacific Islander/Asian-American, 5=Native American=5, 6=Other); dec_o (decision of partner the night of event); attr_o (rating by partner the night of the event, for all 6 attributes); age; field: field of study; field_cd: field code (1=Business/Law, 2= Math/Statistics, 3= Psychology/Sociology, 4= Medicine, 5=

Engineering, 6= Classic/Writing, 7= Philosophy/Religion, 8= Economics/Business/Finance, 9= Education, 10= Chemistry, 11= Social Work, 12= Undergrad – GS, 13= Political/International Affair, 14= Film, 15= Art/Theater, 16= Language, 17= Architecture, 18= GSAS and other); goal: The primary goal of participating in this event (1= Seemed like a fun night out, 2= To meet new people, 3= To get a date, 4=Looking for a serious relationship, 5=To say I did it, 6= Other); go out: How often do they go out (not necessarily on dates)? (1=Several times a week=1, 2=Twice a week, 3=Once a week, 4=Twice a month, 5=Once a month, 6=Several times a year, 7=Almost never=7); 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?

Then, dealing with missing data and recognizing data type are very important. After read speed dating data from a CSV file, I use 0 to replace a part of numerical NA and NULL data, omit some data which are not related to my research.

Simple bar plots statistics and analysis

First of all, 6 bar plots will be conducted. They are “numbers of events”, “ages”, “races”, “study fields”, “goals of dating” and “match or not”.



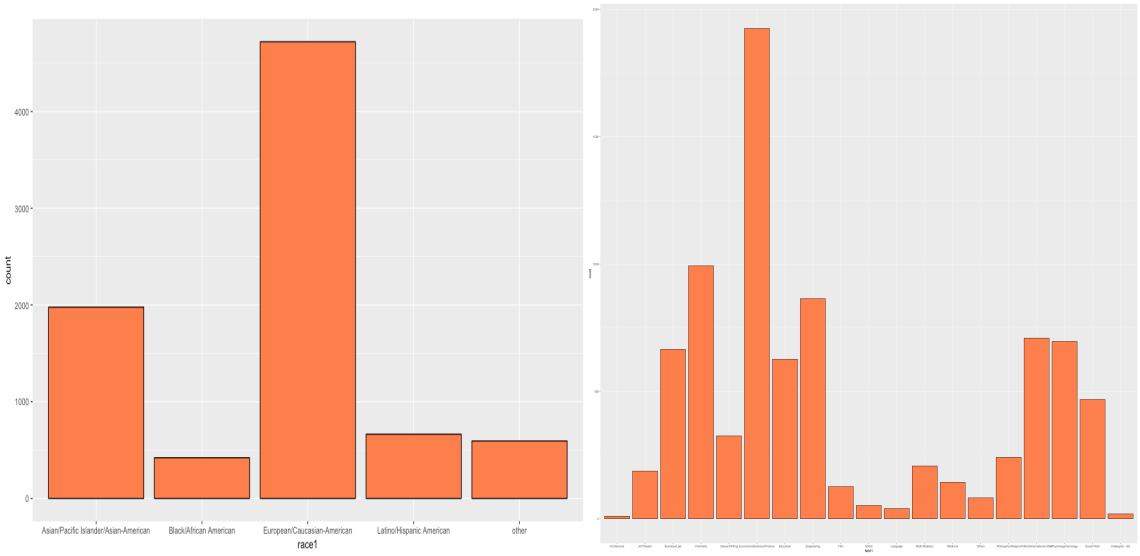
In the “Events” plot, we could find that the more than 20 events, there are nearly 1000 speed dates per event and most of them are over 500 dates per event. It shows that a lot

of people are willing to join this kind of dating.

In the “age” plot, it is obvious to see, the ages of a plenty of people who take part in speed dating are 20-30 years old. Also, I calculated the mean of age, which is around 26-year-old.

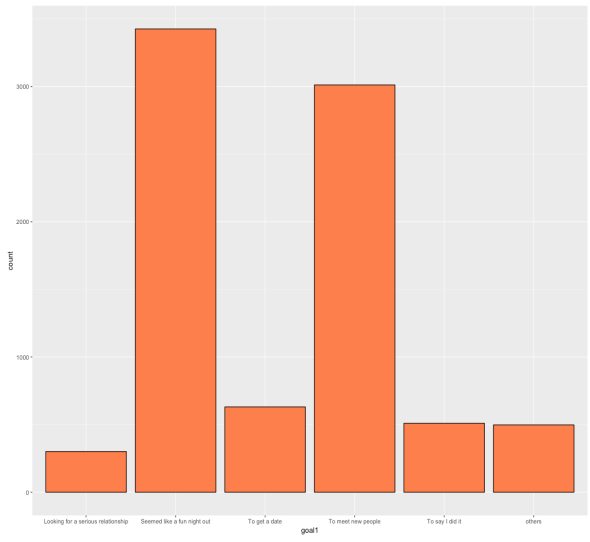
```
> mean(sdd$age)
[1] 26.06004
```

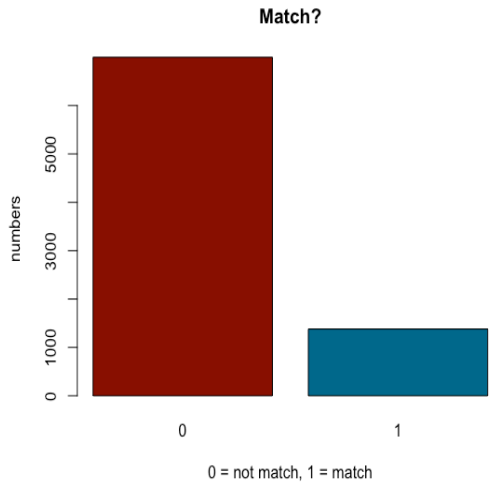
It indicates that one of purpose of taking part in speed dating is to seek their marriage partner at that age.



In the “race” plot, European/Caucasian-American extremely like to take part in speed dating, the numbers of people more than 5000. I consider that the reason could be they are more open-mind in this occasion, enjoy making friends and recognizing people. However, Black/African American rarely participate in speed dating, I think the reason is that they pay more attention to other field, like work, study, etc.

In the “study field” plot, nearly 2000 people who participate in speed dating engage in economics, business or finance field. The least of them work in architecture. Therefore, I consider that people who work in economics, business and finance field are good with people, prefer making friends and social. While, there are few chances for people who work in architecture taking parties and joining this kind of dating, they always like isolate lifestyle.

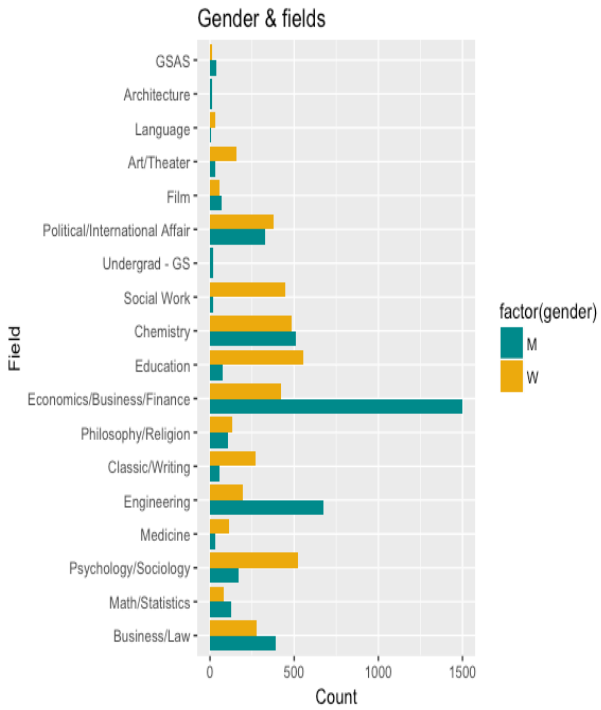




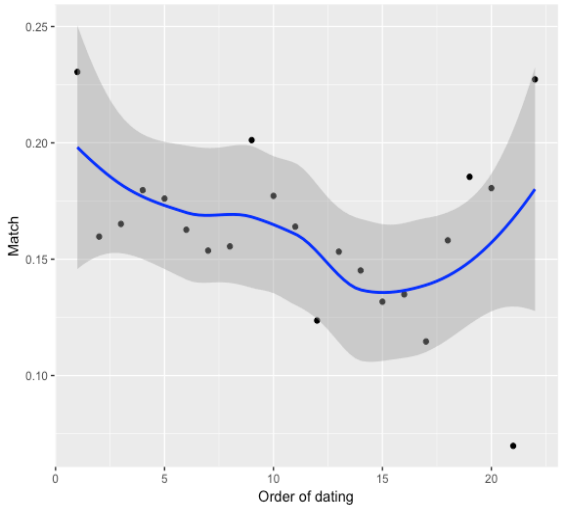
In the “goals of dating” plot, the most people think speed dating seems a fun night out and to meet new people. However, there is a tricky thing, the least of them to look for a serious relationship. Based on this phenomenon, the numbers of people who feel matched, in other words, they are willing to continue with the partner is 1/5 of people who feel not matched.

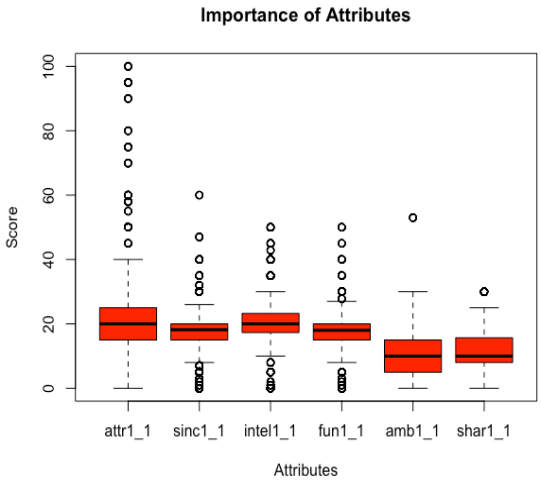
Relationship

Moreover, I will analyze and compare different variables in this part, in order to find relationship among them. Using “ggplot” to draw two bar plots, which are “gender and field” and “gender and race”. Then using “geom_smooth” and points to illustrate relationship between order of dating and match. Additionally, using “boxplot” to analyze the influence of 6 attributes per 100 points for making next decision. The influence of attributes that I would like to do regression and correlation for deeper analysis in the next part.



In the “Gender and fields” graph, it is not hard to find the number of males in Economics, business and finance are the most. Nevertheless, females in education field is the most. In the “Race and Gender” graph, the largest number of men and women are European and Caucasian-American. And none of them are native American. According to these two graphs, we could observe different numbers between males and females in speed dating.

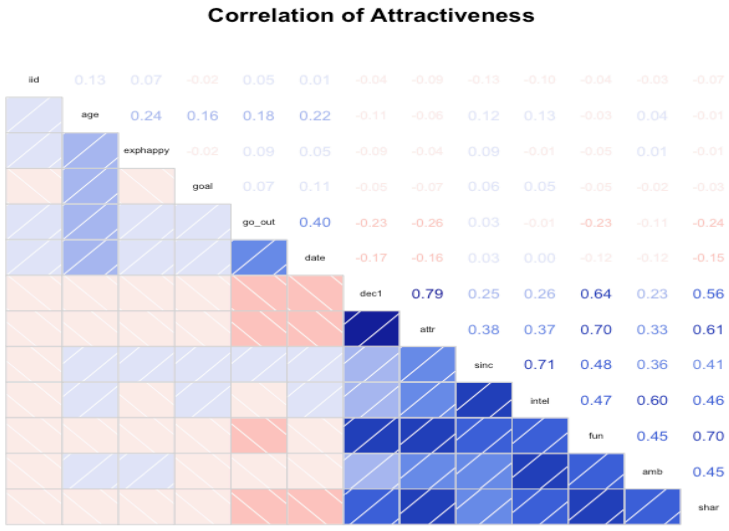




In the “order of dating” graph, I use the mean of “match” and “order” of dating to explore the trend of chances to decide whether to keep dating in the future. We can see at the beginning of data, there is highly possibility to match, however with growing of the number of date for a night when people meet partner, the match rate decreasing gradually till 15 partners, people could rekindle their interests or lose their interests totally, but few of people could insist.

In the box plot, we can find easily, influence of attractiveness is the largest, fun and intelligence followed. Sincere, ambitious and share interests are lower than the others. This result may imply people are superficial, they always focus on attractiveness. However, it is an evidence to reveal the reason why people are willing to take part in the speed dating.

Decision Correlation (6 attributes)



In this correlation, I use subject number (iid), age, gender, income, score of happy from 1-10 (exphappy), goal, how often do they go out(go_out), date and 6 attributes (attractive, sincere, intelligence, fun, ambitious and share interests) to illustrate the correlation among these data. As the correlation graph shows, these 6 attributes have a strong correlation among these variables, it means there is a strong influence of decision,

especially attractiveness. Except these 6 attributes, age is an influence element as well.

Linear Regression

In this part, first, I am going to analyze R-squares, coefficients P-value and residuals of each attribute to find the strongest linear indicators between decision and 6 attributes. Furthermore, I will conduct graphs of 3 strongest linear relation.

- Test 1 (attractiveness, fun, sincere, intelligence, ambitious, share interests)

```
Call:
lm(formula = dec1 ~ attr, data = attr1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46460 -0.09918 -0.00644  0.09618  0.46395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.518879   0.031774  -16.33   <2e-16 ***
attr         0.155720   0.005139   30.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1467 on 549 degrees of freedom
Multiple R-squared:  0.6258,    Adjusted R-squared:  0.6251
F-statistic: 918.2 on 1 and 549 DF,  p-value: < 2.2e-16
```

decision and attractiveness

```
Call:
lm(formula = dec1 ~ fun, data = attr1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6369 -0.1291 -0.0019  0.1284  0.6240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.454360   0.046065  -9.863   <2e-16 ***
fun          0.142343   0.007346  19.377   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1848 on 549 degrees of freedom
Multiple R-squared:  0.4061,    Adjusted R-squared:  0.4051
F-statistic: 375.5 on 1 and 549 DF,  p-value: < 2.2e-16
```

decision and fun

```
Call:
lm(formula = dec1 ~ sinc, data = attr1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50122 -0.17622 -0.00967  0.16837  0.71025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09587    0.08527  -1.124    0.261
sinc         0.07464    0.01213   6.152 1.48e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.232 on 549 degrees of freedom
Multiple R-squared:  0.06448,   Adjusted R-squared:  0.06278
F-statistic: 37.84 on 1 and 549 DF,  p-value: 1.479e-09
```

decision and sincere

```
Call:
lm(formula = dec1 ~ intel, data = attr1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51645 -0.17499 -0.00799  0.17110  0.69098

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13736    0.08855  -1.551    0.121
intel        0.07877    0.01232   6.392 3.51e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2314 on 549 degrees of freedom
Multiple R-squared:  0.06926,   Adjusted R-squared:  0.06756
F-statistic: 40.85 on 1 and 549 DF,  p-value: 3.509e-10
```

decision and intelligence

```
Call:
lm(formula = dec1 ~ amb, data = attr1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53276 -0.18242 -0.00955  0.17067  0.56409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05752    0.06713   0.857    0.392
amb          0.05867    0.01060   5.537 4.78e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2334 on 549 degrees of freedom
Multiple R-squared:  0.05289,   Adjusted R-squared:  0.05116
F-statistic: 30.66 on 1 and 549 DF,  p-value: 4.777e-08
```

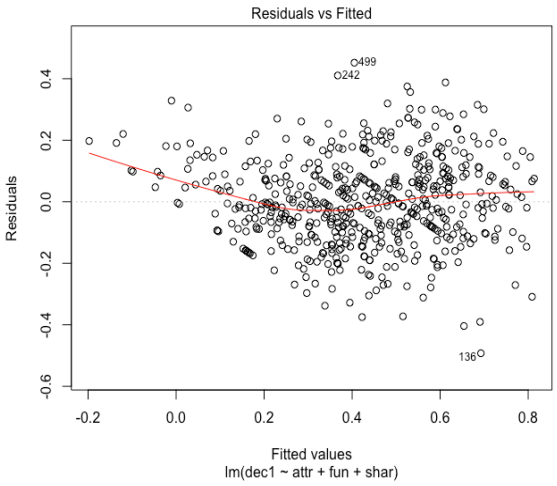
decision and ambitious

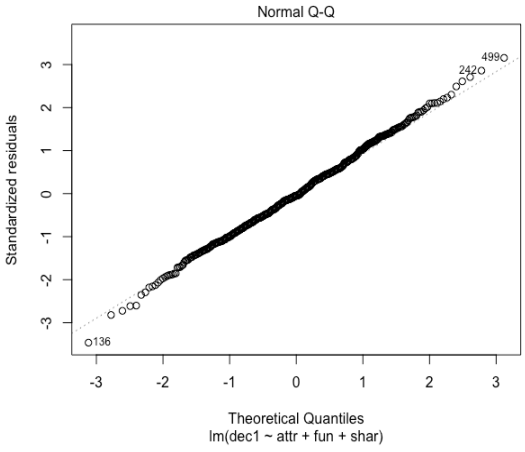
```
Call:
lm(formula = dec1 ~ shar, data = attr1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59987 -0.13881 -0.00744  0.12851  0.60733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.230734   0.041781  -5.522 5.17e-08 ***
shar         0.135606   0.008461  16.027  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.198 on 549 degrees of freedom
Multiple R-squared:  0.3187,    Adjusted R-squared:  0.3175
F-statistic: 256.9 on 1 and 549 DF,  p-value: < 2.2e-16
```

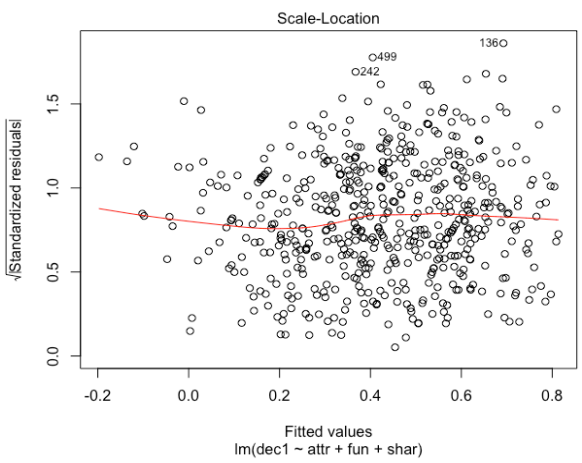
decision and share interest

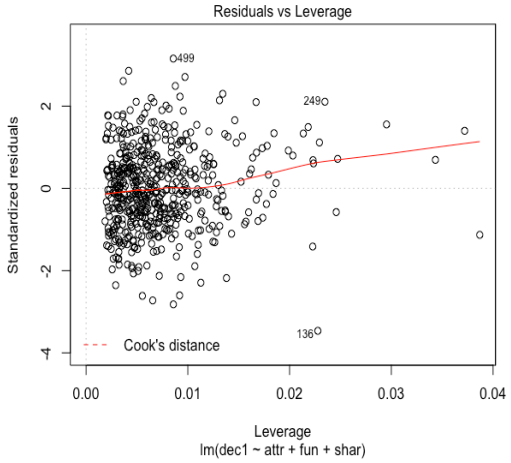
According to these results of test, we can see in the first test, R-square of decision and attractiveness is 0.63, residual is 0.15, p-value less than 2.2 e-16, there is large R-square, small error and p-value, which means attractiveness has very tight relationship with decision. Likewise, in the fun and decision, 0.4 R-square, 0.18 residual and p-value less than 2.2 e-16. Share interest and decision have 0. 32 R-square, 0.198 residual and p-value less than 2.2 e-16. So, attractiveness, fun and share interest are the most 3 influential attributes, additionally, I use these 3 attributes make a regression analysis.

- Test 2 (attr+fun+shar)



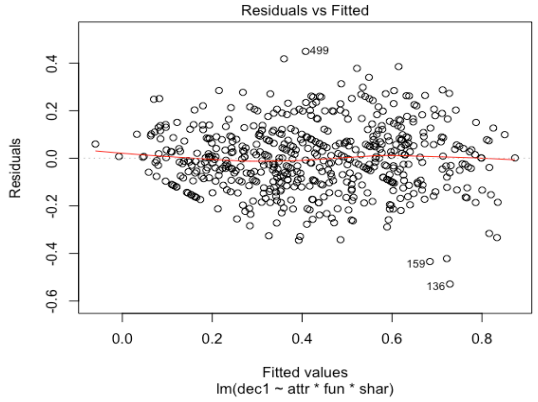


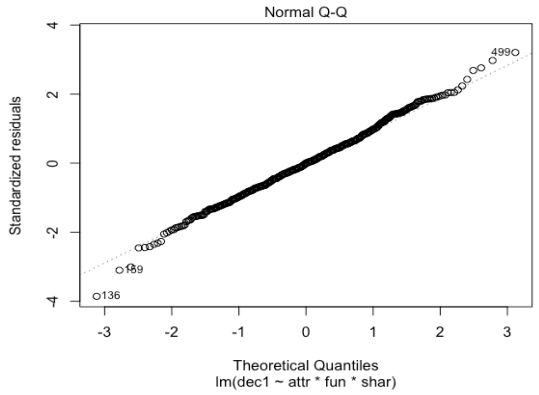


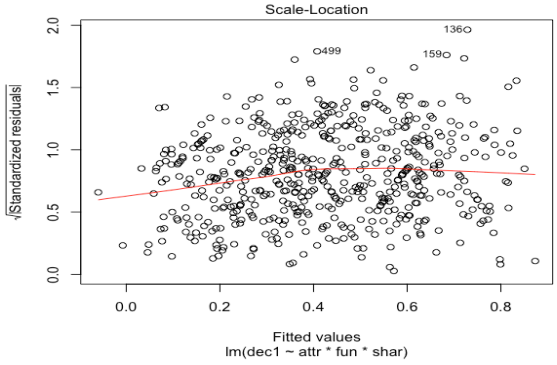


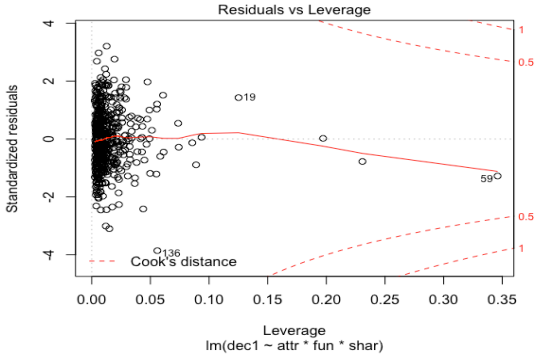
At first, I calculate relationship between decision and attractiveness, fun and share interests, I find that only few of data effect result of regression model, most of them fit well. However, we can see in the residual and fitted graph and leverage graph are not flat , so we need to adjust a little.

- Test 3 (attr*fun*shar)



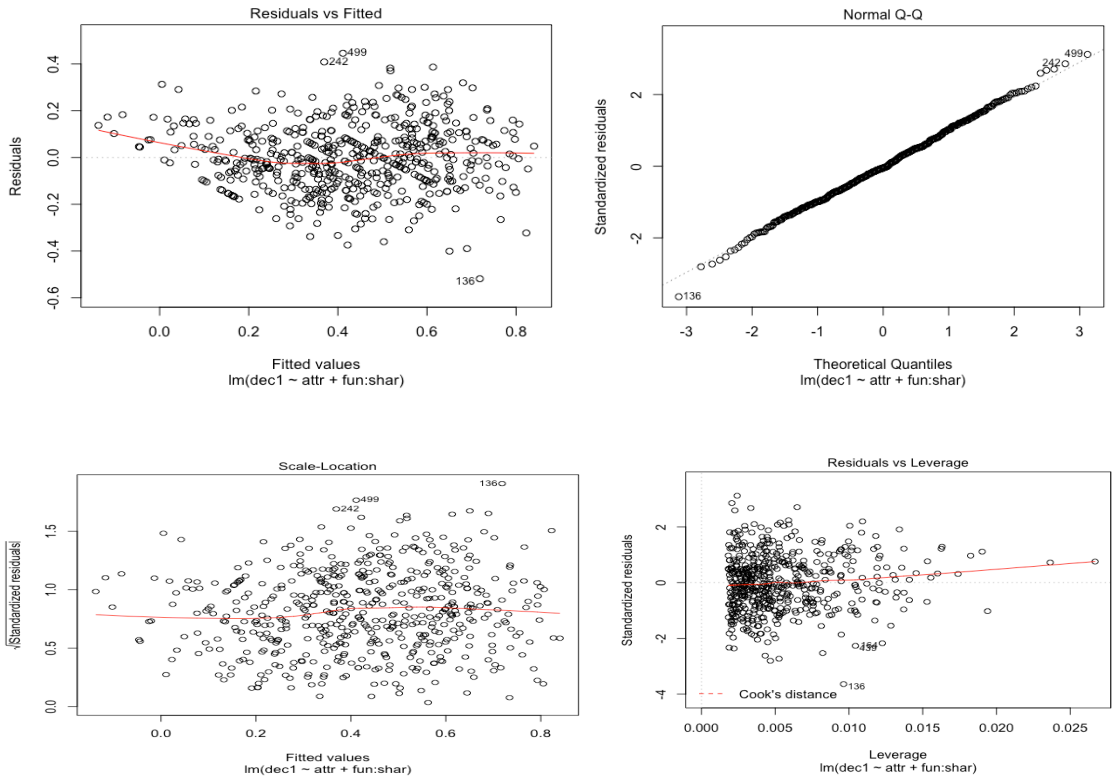






Through another computation by using multiplication among attractiveness, fun and share interests, we can find the regression graphs are fitted better than using addition.

- Test 4 (attr+fun:shar)



According to this fitting way, reducing the effect of residuals, the regression is better than test 2 and 3.

Conclusion

By using different plots, for instance, bar plot, box plot, correlation matrix and linear regression, I calculate determinants whether people decide to next dating and pick up three strongest influential factors of their decision, which are attractiveness, fun and share interests. In addition, this project explore relationship between two variables, such as gender and field, order and match.

In the future research, I'd like to analyze more attributes how to affect match rate and how to improve the effectiveness of speed dating. For this project, I'm willing to use more visuals to represent diverse and complete results.

Script

```
###set up

library(readr)
library(dplyr)
library(tibble)
library(stringr)
library(ggplot2)
library(grid)
library(scales)
library(reshape2)
library(hexbin)
library(ellipse)
library(corrplot)
library(lattice)
library(MASS)
library(car)
library(splines)
library(corrgram)

### read dataset
sdd<- read.csv("Untitled_Message/Speed Dating Data.csv", header = T, stringsAsFactors =
F)
dim(sdd)
str(sdd)
sum(is.na(sdd))

###check missing data
```

```
missing.data<- c("NA", "")
missing.data<-0
sdd[is.na(sdd)]<-0
sum(is.na(sdd))
View(sdd)
colnames(sdd)



##### bar plots analysis #####

###Events
qplot(data=sdd, wave, binwidth=1, colour=I("black"), fill=I("coral"),
        xlab="Events", ylab="numbers")

###ages
str(sdd$age)
mean(sdd$age)
qplot(data=sdd, age, binwidth=1.5, colour=I("black"), fill=I("coral"))

###races
sddr<-sdd %>% mutate(race1=ifelse(race_o %in% 1,"Black/African American",
                                  ifelse(race_o %in% 2, "European/Caucasian-American",
                                  ifelse(race_o %in% 3,"Latino/Hispanic American",
                                  ifelse(race_o %in% 4, "Asian/Pacific Islander/Asian-
American",
                                  ifelse(race_o %in% 5, "Native American",
                                  "other"))))))
qplot(data=sddr, x=race1, geom = "bar", colour=I("black"), fill=I("coral"))

###fields
```

```
sdd$field_cd=factor(sdd$field_cd)
sddf<-sdd %>% mutate(field1=ifelse(field_cd %in% 1,"Business/Law",
                                   ifelse(field_cd %in% 2, "Math/Statistics",
                                          ifelse(field_cd %in%
3,"Psychology/Sociology",
                                                 ifelse(field_cd %in% 4,
"Medicine",
                                                        ifelse(field_cd %in% 5,
"Engineering",

ifelse(field_cd %in% 6,"Classic/Writing",

ifelse(field_cd %in% 7, "Philosophy/Religion",

ifelse(field_cd %in% 8,"Economics/Business/Finance",

ifelse(field_cd %in% 9, "Education",

ifelse(field_cd %in% 10, "Chemistry",

ifelse(field_cd %in% 11,"Social Work",

ifelse(field_cd %in% 12, "Undergrad - GS",

ifelse(field_cd %in% 13,"Political/International Affair",

ifelse(field_cd %in% 14, "Film",

ifelse(field_cd %in% 15, "Art/Theater",
```

```
ifelse(field_cd %in% 16, "Language",

ifelse(field_cd %in% 17,"Architecture",

ifelse(field_cd %in% 18,"GSAS","Others"))))))))))))))))))

qplot(data=sddf,    x=field1, geom = "bar",colour=I("black") , fill=I("coral"))

###goal

sddg<-sdd %>% mutate(goal1=ifelse(goal %in% 1,"Seemed like a fun night out",
                                  ifelse(goal %in% 2, "To meet new people",
                                         ifelse(goal %in% 3,"To get a date",
                                                ifelse(goal %in% 4, "Looking
for a serious relationship",
                                                       ifelse(goal  %in%  5,
"To say I did it",

"others"))))))

qplot(data=sddg,    x=goal1, geom = "bar",colour=I("black") , fill=I("coral"))

###match
barplot(
   table(sdd$match),xlab = '0 = not match, 1 = match', ylab = 'numbers',
   main = "Match?",
   col = c("darkred","deepskyblue4"))

##### Relationship #####
```

```
### gender and fields

sdd[sdd$gender == 0,]$gender <- "W"
sdd[sdd$gender == 1,]$gender <- "M"
field2 <- sdd[!is.na(sdd$field_cd),] %>%
   group_by(gender, field_cd) %>%
   summarise(n = n())


field3<-c("Business/Law","Math/Statistics","Psychology/Sociology","Medicine",

"Engineering","Classic/Writing","Philosophy/Religion","Economics/Business/Finance",
                "Education","Chemistry","Social                Work","Undergrad                -
GS","Political/International Affair",
                "Film","Art/Theater","Language","Architecture","GSAS","Others")

ggplot(field2, aes(x = field_cd, y=n, fill = factor(gender))) +
   geom_bar(stat = "identity", position = "dodge") +
   scale_fill_discrete(name = "Gender") +
   xlab("Field") + ylab("Count") + ggtitle("Gender & fields") +
   scale_x_discrete(labels = field3, breaks = 1:19) +
   coord_flip()+
   scale_fill_manual(values=c("darkcyan","darkgoldenrod2"))

###gender and age

race2 <- c(
   "Black/African American",
   "European/Caucasian-American",
   "Latino/Hispanic American",
```

```
    "Asian/Pacific Islander/Asian-American",
    "Native American",
    "Other"
)

race3 <- sdd[!is.na(sdd$race),] %>%
    group_by(gender, race) %>%
    summarise(
        n = n()
    )

ggplot(race3, aes(x = race, y = n, fill = factor(gender))) +
    geom_bar(stat = "identity", position = "dodge") +
    scale_fill_discrete(name = "Gender") +
    xlab("Race") + ylab("Count") + ggtitle("Race & Gender") +
    scale_x_continuous(labels = race2, breaks = 1:6) +
    coord_flip()+
    scale_fill_manual(values=c("darkcyan","darkgoldenrod2"))

### order and match
om<-sdd %>%
    group_by(order) %>%
    summarise(average=mean(match, na.rm=TRUE))
qplot(data=om, x=order, y=average, xlab="Order of dating", ylab="Match") +
geom_smooth(method=loess,color=I("blue"))

###points of attributes
sdd <- sdd %>%
    mutate(sum1_1=attr1_1+sinc1_1+intel1_1+fun1_1+amb1_1+shar1_1) %>%
    mutate(attr1_1n=attr1_1/(sum1_1/100)) %>%
```

```
   mutate(sinc1_1n=sinc1_1/(sum1_1/100)) %>%
   mutate(intel1_1n=intel1_1/(sum1_1/100)) %>%
   mutate(amb1_1n=amb1_1/(sum1_1/100)) %>%
   mutate(shar1_1n=shar1_1/(sum1_1/100))

boxplot(sdd[,65:70], col="red",
            xlab='Attributes',ylab='Score',main='Importance of Attributes')



##### decision correlation (6 attributes) #####

attr1 <- sdd %>% group_by(iid, age, gender, income, exphappy, goal, go_out, date) %>%
   summarise(dec1= mean(dec_o, na.rm = T),
                  attr = mean(attr_o, na.rm = T),
                  sinc = mean(sinc_o, na.rm = T),
                  intel = mean(intel_o, na.rm = T),
                  fun = mean(fun_o, na.rm = T),
                  amb = mean(amb_o, na.rm = T),
                  shar = mean(shar_o, na.rm = T))



quartz()
corrgram(attr1,
              upper.panel=panel.cor, text.panel=panel.txt,
              main="Correlation of Attractiveness")

##### Linear Regression #####

###test 1.1: decision~attractive
lm.attr11 = lm(dec1~attr,data=attr1)
```

```
summary(lm.attr11)

###test 1.2: decision~fun
lm.attr12 = lm(dec1~fun,data=attr1)
summary(lm.attr12)



###test 1.3: decision~sincere
lm.attr13 = lm(dec1~sinc,data=attr1)
summary(lm.attr13)

###test 1.4: decision~intelligece
lm.attr14 = lm(dec1~intel,data=attr1)
summary(lm.attr14)

###test 1.5: decision~ambitious
lm.attr15 = lm(dec1~amb,data=attr1)
summary(lm.attr15)

###test 1.6: decision~share interests
lm.attr16 = lm(dec1~shar,data=attr1)
summary(lm.attr16)

##### Choose 3 strong variables #####
### attractive,fun,shar(R-squared > 0.3)

### test 2.1
lm.attr21 = lm(dec1~attr+fun+shar,data=attr1)
summary(lm.attr21)
plot(lm.attr21)
```

```
### test 2.2
lm.attr22 = lm(dec1~attr*fun*shar,data=attr1)
summary(lm.attr22)
plot(lm.attr22)

### test 2.3
lm.attr23 = lm(dec1~attr+fun:shar,data=attr1)
summary(lm.attr23)
plot(lm.attr23)
```