1. SPEED DATING EXPERIMENT DESIGN

The dataset we used comes from an experiment conducted at Columbia University by Ray Fisman and Sheena Iyengar, both of whom are faculty at Columbia Business School. The data was gathered over three years from dating events between 2002-04.

Below is the experiment design:

- The experiment is limited to heterosexual dating.
- As a precursor to the event, participants fill out a survey providing their personal information like Age, Career Field, Income and Race.
- The experiment is split into waves. Each wave has 10 females and 10 males.
- During a wave, each participant "dates" every participant of the opposite gender.
- Each date lasts for 4 minutes. (The short duration of the date makes the experiment reliable to analyse 'First Impressions' and 'Love at First Sight')
- At the end of their four minutes, participants rated their date on different attributes.
- There were also follow up dates once a couple decides they want to take it further, but, the team restricted its analysis to 'First Dates' because we seek to improve our online match recommendation engine for people who most likely haven't dated each other previously.

2. QUESTIONS OF INTEREST

We framed questions that we believed will help us come closest to understanding the extremely complex science of romance. The questions progress from providing descriptive and inferential analysis of Speed Dating data through various sophisticated statistical techniques. Finally, we arrive at a predictive model for expected “like” rating received by an individual taking into account all key factors.

Below are the ‘Questions of Interest’ which were analysed in the project:

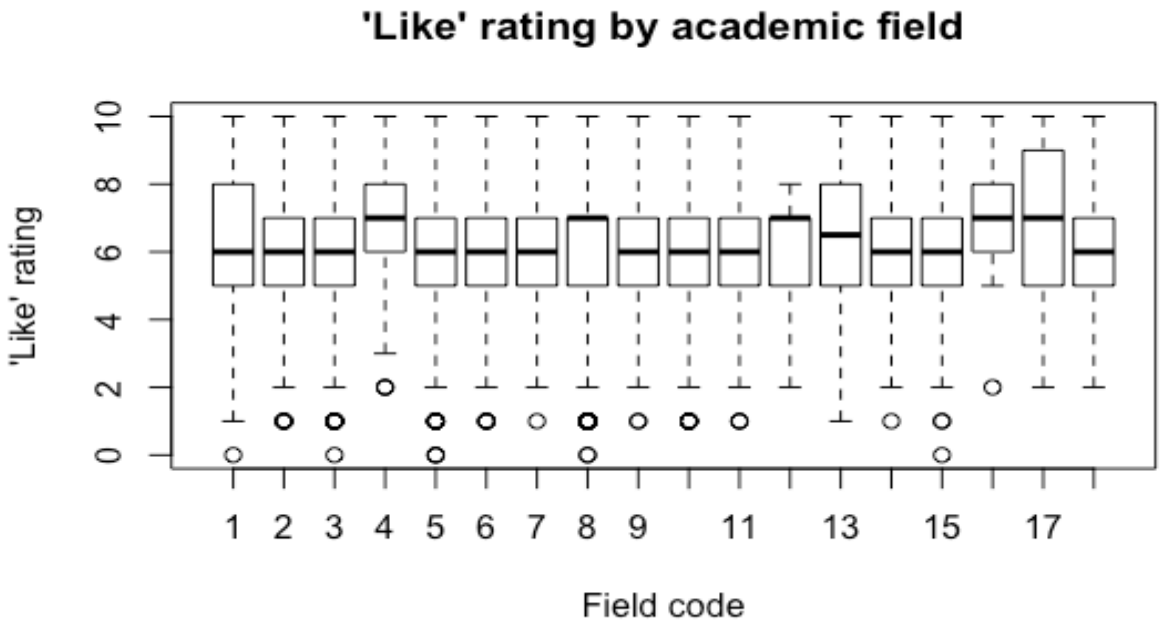
1. Is a person’s academic field associated with how “liked” they are by a prospective partner?
2. Do opposites attract or birds of a feather flock together?
3. How is age associated with the possibility of finding a partner for males? For females?
4. a) What’s the gender difference in mate selection? – Attributes Difference

   b) Are women more selective then men?
5. What influences love at first sight?

3. ANALYSIS

4.1 Is a person's academic field associated with how "liked" they are by a prospective partner?

We conducted an analysis to examine whether or not there is a relationship between the subject's field of study and the partner's "like" rating of the subject.



Field code key:

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

Result –

The results of our analysis support the hypothesis that there is a relationship between field and 'like' rating.

We then followed up on these results to see which fields had significant relationships with each other when it came to their associated "like" level. We were willing to accept a 5% margin of error for this analysis, and found significant relationships between 21 different pairs of fields.

In general, medicine and engineering were the fields that had the highest number of significant relationships with other fields in terms of "like" rating. All medicine "like" ratings were higher than other fields with which it showed a significant relationship: law, math, social science, engineering, English, education, bio/chem/physics, social work, and fine arts. All engineering "like" ratings were lower than the other fields with which it showed a significant relationship: law, social science, medicine, history, business, education, bio/chem/physics, social work, political science, and languages. Math also showed a significantly lower average like rating than law, medicine, business, political science, and languages, with which it exhibited significant relationships.

The 'like' rating is based on a scale of 0 to 10, with 0 being "not at all" and 10 being "like very much". The largest differences in average "like" rating were seen between medicine and math (medicine 1.53 greater), languages and math (languages 1.22 greater), medicine and engineering (medicine 1.11 greater), and languages and engineering (languages 1.17 greater).

Conclusions –

Our analysis shows that subjects studying medicine received significantly higher like ratings than a number of other fields, while those studying engineering and math received significantly lower like ratings. A follow up project could attempt to examine why this is so.

4.2 Do opposites attract or birds of a feather flock together?

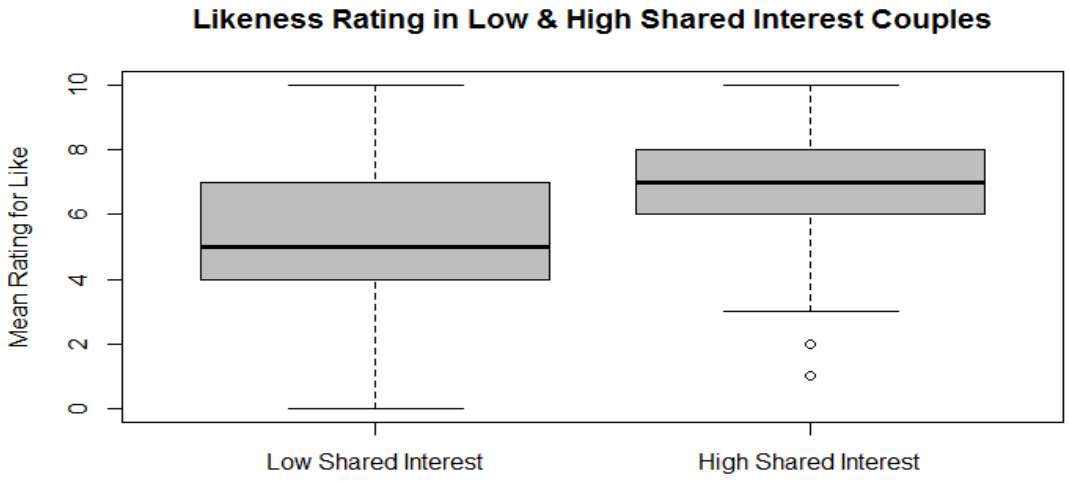
We wanted to put to test the popular saying that 'Opposites attract each other'. In order to do this, we split the dataset into two subsets by the median Shared Interest Level. Then we compared the 'Mean Like Rating' received by subject in the Low Shared Interest Couples with the High Shared Interest Couples.

*Subject – The individual receiving the rating

*Partner – The individual rating the Subject

Below are the Mean Like Ratings we saw in the two groups:

	Low Shared Interest	High Shared Interest
Mean Like Rating	5.277	6.836



Result –

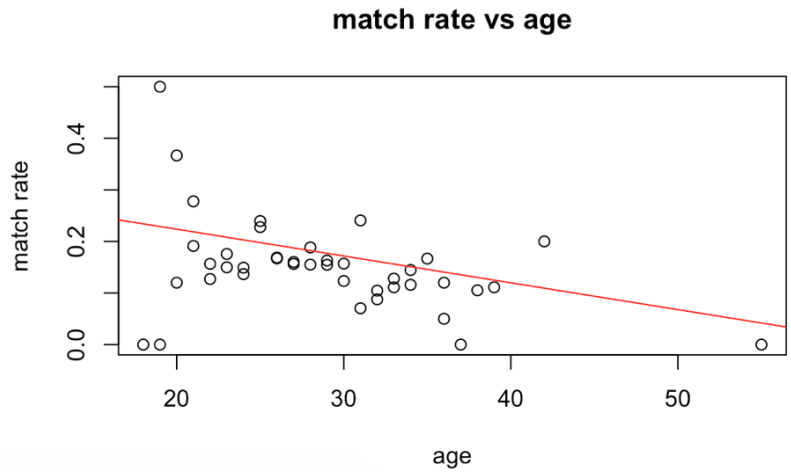
We found that there is a statistically significant difference between the Mean Like Ratings in 'High Shared Interest Couples' vs 'Low Shared Interest Couples'.

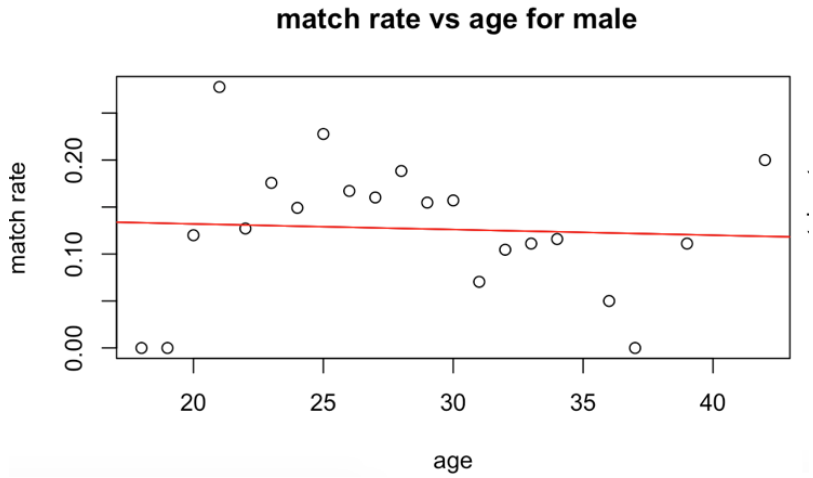
Conclusion –

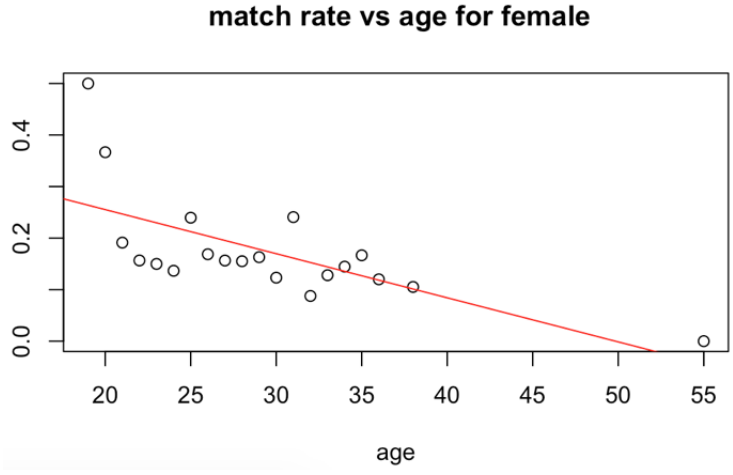
On first dates, couples with high shared interest hit it off more than others.

4.3 How does age influence the possibility of finding a partner for males? For females?

We wanted to understand how does age affect the match rate for both male and female. What we did is group the mean match rate by gender and age, and then built two regression model for male and female. Here is the graph of our result:







Result –

The result shows that for our entire sample (include both male and female), the overall trend for match rate as age increasing is downward as we expect.

However, our coefficient estimates show that there is no significant relationship between age and match rate for male at all. The same conclusion can also be drawn from the graph list above.

As for female, we found a significant age estimate for match rate. The result indicates that the match rate will be brought down by 8.5% with one-year increase in age.

Conclusion –

Our analysis indicates that age does affect the match rate, but only has an significant influence for female, as they might lose attractiveness for male as time goes by. But the effect disappear when it comes to male.

4.4 What's the gender difference in mate selection? Are women more selective then men?

We performed a statistical analysis to explore how men and women's preferences differ in terms of dating. We analysed how each person rated their partner on six attributes (on a 0-10 scale): physical attractiveness, intelligence, sincerity, fun, shared interests, and ambition, and then looked at the relationship of these attributes with overall "like" rating for each gender. We explored the speed dating data focusing on gender difference and found that men and women have different concerns in mate selection. Here is the model we constructed for gender differences in attribute weights:

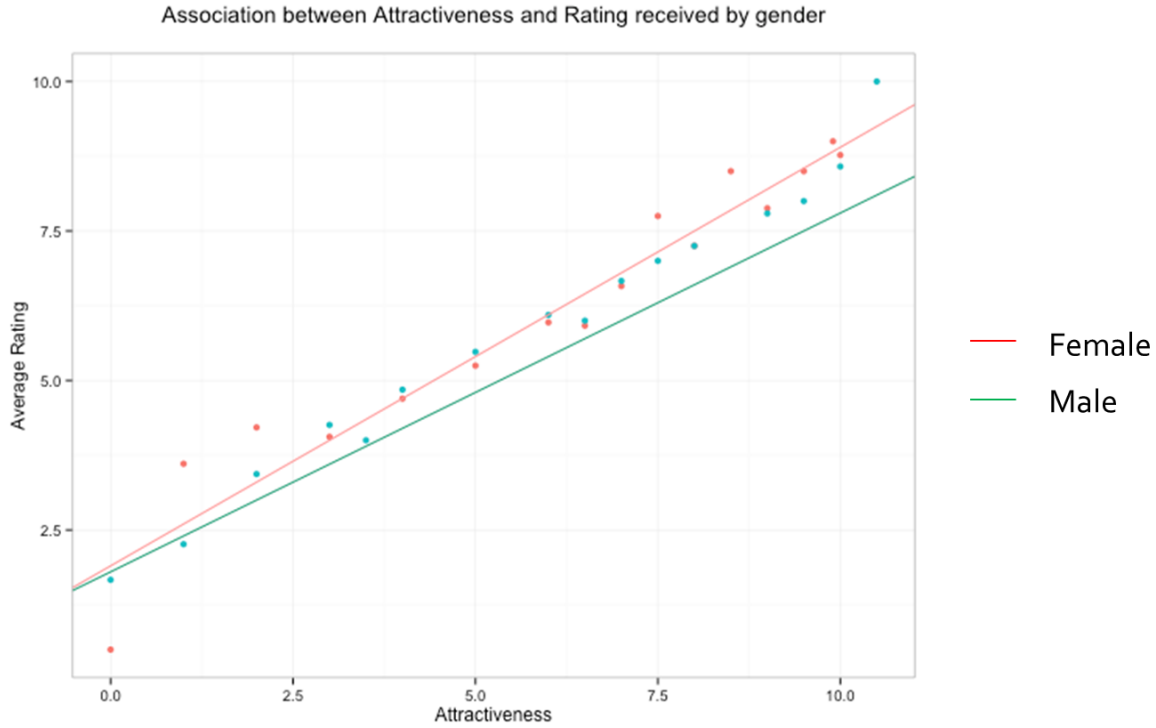
Rating Received = .19 - .37 Male + .33 Attractiveness + .05 Intelligence + .11 Sincere + 0.22 Fun + .25 Shared Interest - 0.07 Male * Attractiveness + 0.10 Male*Intelligence

The model suggests that valuation of attributes by men and women are different. *Women put greater weight on intelligence than men do, while men place more value on physical appearance*. The magnitudes of these differences are relatively large. On average, each additional attractiveness point (on a 10-point scale) increases rating received by female by 0.07 more than it increases the rating male received. The implied effect of intelligence on rating received is 0.15 for male compared with only 0.05 for female.

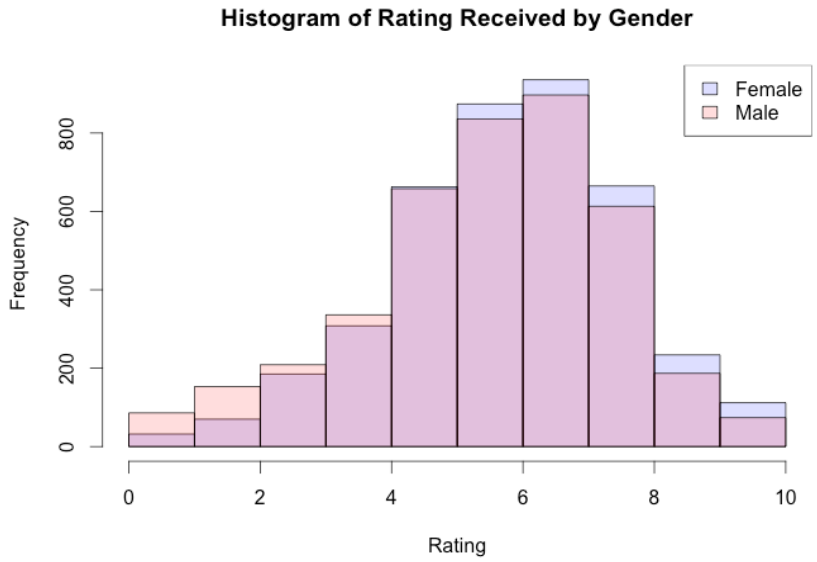
Result –

Physical attraction and intelligence had different levels of association with "like" rating for men and women. Overall, our results show that women value intelligence more than men do, while men put more weight on physical appearance.

Additionally, we analysed the overall relationship in "like" ratings between men and women. In general, men tended to give women higher "like" ratings (8, 9, 10), while women tended to give more low "like" ratings to men (0, 1, 2). On average, ratings received by men were 0.28 lower than ratings received by women. This supports the hypothesis that women are more selective in their dating preferences.


Association between Attractiveness and Rating received by gender
Average Rating
10.0
7.5
5.0
2.5
0.0
2.5
5.0
7.5
10.0
Attractiveness
Female
Male


Association between Intelligence and Rating received by gender
Average Rating
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
Intelligence
Female
Male



Conclusion –

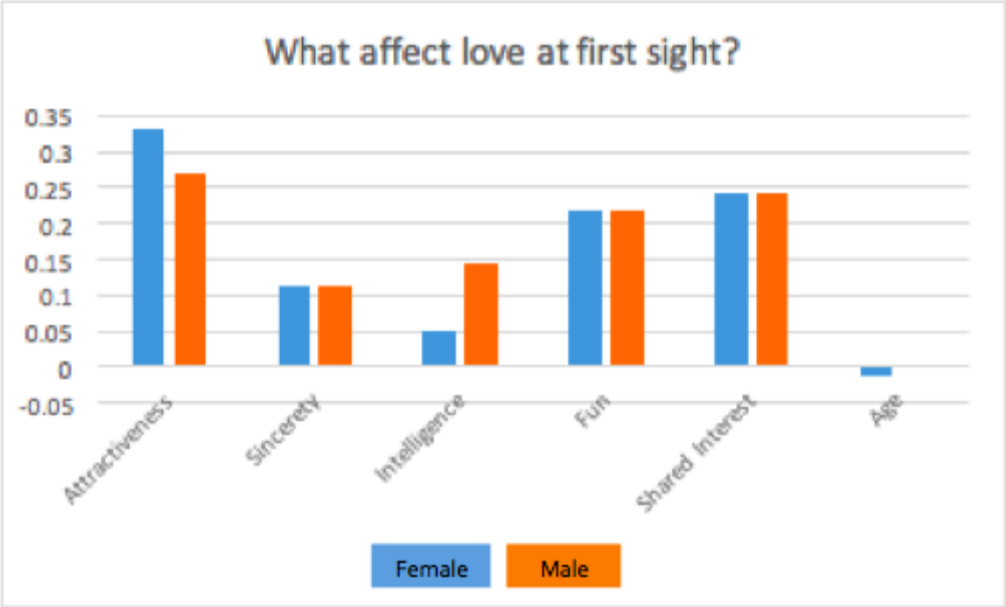
Women value intelligence more than men do, while men put more weight on physical appearance. Women are also more selective when it comes to their dating partners.

4.5 What influences love at first sight?

We performed a statistical analysis to explore what attributes are associated with higher "like" ratings. We analysed how each person rated their partner on six attributes (on a 0-10 scale): physical attractiveness, intelligence, sincerity, fun, shared interests, and ambition, and then looked at the relationship of these attributes with overall "like" rating for each gender. We also incorporated the data on age and match rate (see section 4.3).

Result –

All attributes showed a significant relationship with "like" rating, except ambition, which did not seem to be associated with "like" rating. Sincerity, fun, and shared interests had equal levels of association with "like" rating for men and women, while physical attraction and intelligence had different levels of association with "like" rating for men and women.



Conclusion –

Sincerity, fun, and shared interests have a significant relationship with "like" rating for both men and women, while ambition does not appear to have a relationship with "like" rating. Intelligence is a trait that women value more than men, while men put more weight on physical appearance in women.

Footnote –

[1] The Causality relationship is built up by professors who conduct original studies through psychological theory. See Reference and Appendix for more information.

4. ALL CONCLUSIONS

- **Is a person's academic field associated with how "liked" they are by a prospective partner?**
    - Our analysis shows that subjects studying medicine received significantly higher like ratings than a number of other fields, while those studying engineering and math received significantly lower like ratings. A follow up project could attempt to examine why this is so.

- **Do opposites attract or birds of a feather flock together?**
    - 'Opposites Attract!' seems to be a myth at least for the first impressions. It is the 'High Shared Interest Couples' who hit it off better. Though, it should be mentioned that the analysis is limited to a speed dating experiment when couples get only 4 minutes to judge each other and the results could be very different in the follow up dates over a long period.

- **How is age associated with the possibility of finding a partner for males? For females?**
    - Our analysis indicates that age does affect the match rate, but only has a significant influence for females, as they might lose attractiveness for males as time goes by. But the effect disappears when it comes to males.

- **What is the gender difference in mate selection?**
    - Women value intelligence more than men do, while men put more weight on physical appearance. Women are also more selective when it comes to their dating partners

- **What influences love at first sight?**
    - Sincerity, fun, and shared interests have a significant relationship with "like" rating for both men and women, while ambition does not appear to have a relationship with "like" rating. Intelligence is a trait that women value more than men, while men put more weight on physical appearance in women.

5. LIMITATIONS

Overall limitations for the study:

- Random sampling
  - The subjects of the study were all students (undergraduate, graduate, or professional school) at Columbia University. This analysis is robust if the population under study is students at an elite private urban university. However, the conclusions likely cannot be applied to other populations or the general American population.

- Subjectivity
  - Subjects were asked to rate their partners on a numerical scale. However, perceptions of beauty, intelligence, and love can vary greatly from one person to another, and are not easily quantifiable. This may have introduced some additional error into the study.

- Treatment of ordinal data as continuous
  - We treated the 0-10 "like" rating, an ordinal level variable, as continuous for our analysis. However, according to Menard (2002): "When the dependent variable is ordinal but has five or more categories, treating the variable as though it were continuous is acceptable."

Limitations for the questions addressed:

1. Is a person's academic field associated with how "liked" they are by a prospective partner?
   a. Small sample size for "languages" category: only 40 data points were recorded, which means four language majors were "speed dating" with 10 partners each. Given that only four language majors were being "liked," it is possible that the small sample size has skewed the results.
   b. Additionally, these results are called into question due to issues with the variance assumption for the test (see appendix 8.1).

2. Do opposites attract or birds of a feather flock together?
   No limitations to mention. Number of observations > 30. (see appendix 8.2)

3. How is age associated with the possibility of finding a partner for males? For females?
   a) These results are called into question due to issues with the normality assumption for the test of age and match rate for females. (see appendix 8.3)
   b) There might exist a selection bias as old female tend to be those who are less attractive than other women who find a partner in an early age.

4. What is the gender difference in mate selection?
   No limitations to mention. All assumptions hold good. (see appendix 8.4)

5. What influences love at first sight?

   The Causality relationship is built up by professors who conduct original studies through psychological theory. See Reference and Appendix for more information.

6. APPENDIX

i. *Is a person's academic field associated with how "liked" they are by a prospective partner?*

Statistical tests performed: **ANOVA**, **TukeyHSD**

H_0: No relationship between field (predictor) and "like" rating (response)

H_a: Relationship

Assumptions:

-random sampling

-stability

-for each level i of the predictor variable:

-If any $n_i < 30$, response variable is normally distributed within groups

$-\sigma 1 = \sigma 2 = \ldots = \sigma k$

Data:

Field code	Field	Number of responses (n)	Average "like" rating (0-10)
1	Law	650	6.206
2	Math	202	5.658
3	Social Science, Psychologist	670	6.103
4	Medical Science, Pharmaceuticals, and Bio Tech	138	6.812
5	Engineering	828	5.703
6	English/Creative Writing/ Journalism	317	6.084
7	History/Religion/Philosophy	238	6.172
8	Business/Econ/Finance	1863	6.278
9	Education, Academia	610	6.088
10	Biological Sciences/Chemistry/Physics	978	6.098

11	Social Work	461	6.084
12	Undergrad/undecided	19	6.158
13	Political Science/International Affairs	689	6.335
14	Film	122	6.111
15	Fine Arts/Arts Administration	170	5.959
16	Languages	40	6.875
17	Architecture	10	6.700
18	Other	49	6.082
		Total: 8054	Grand mean: 6.13145

Running the ANOVA:

```
attach(date)
field_cd=factor(field_cd)
fieldanova<-aov(like_o~field_cd)
summary(fieldanova)

##               Df Sum Sq Mean Sq F value   Pr(>F)
## field_cd      17    369  21.684    6.48 1.49e-15 ***
## Residuals   8036  26893   3.347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 324 observations deleted due to missingness
```

P-value of 1.49e-15 indicates a very low chance of getting data as or more extreme than we did, if H0 is true (there is no relationship between field and "like" rating). Therefore, we reject H0 and accept Ha – there is evidence for a relationship.

Testing assumptions:

Variance assumption

H0: $\sigma_1=\sigma_2=\sigma_3=...=\sigma_k$
Ha: At least one $\sigma \neq$ others

Levene Test

```
leveneTest(like_o,field_cd)

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value   Pr(>F)
## group   17  2.1062 0.004938 **
##       8036
```

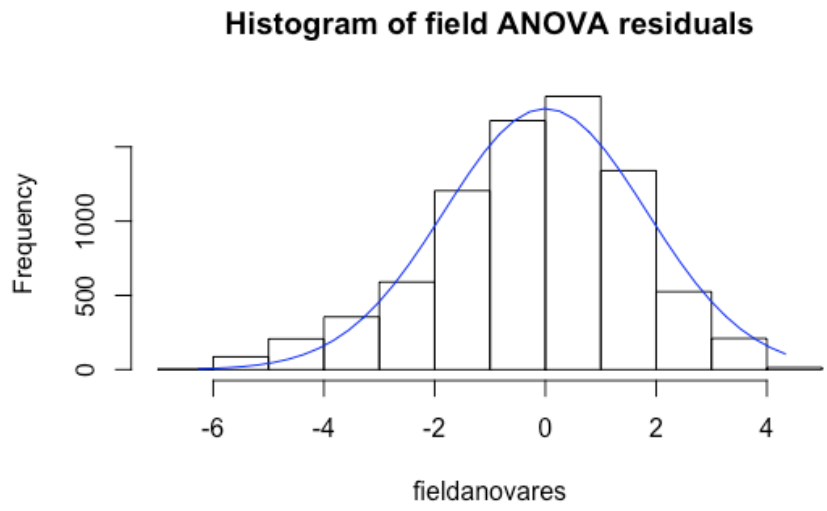
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This assumption is problematic. We would need to look into this and potentially adjust the data to address this issue.

Testing normality:

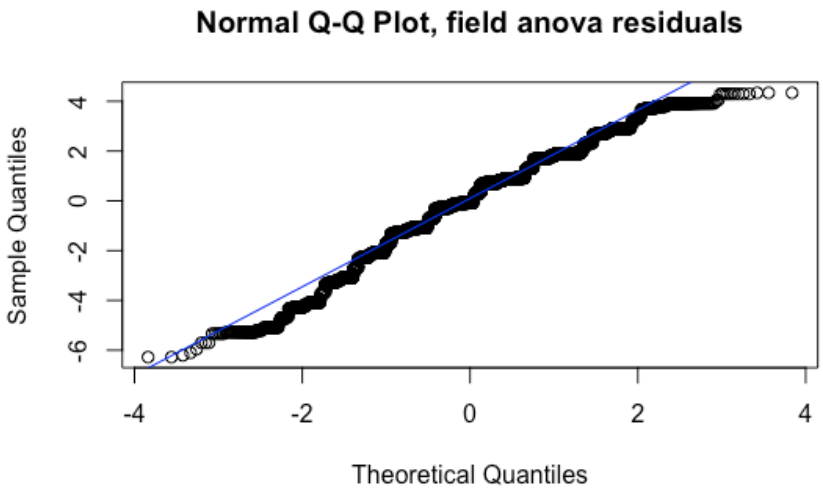
(because $n_i < 30$ for field code 12 (undergrad/undecided) and field code 17 (architecture))

```
fieldanovares<-fieldanova$residuals
h<-hist(fieldanovares, main = "Histogram of field ANOVA residuals")
x<-fieldanovares
xfit<-seq(min(x), max(x), length = 40)
yfit<-dnorm(xfit, mean = mean(x), sd = sd(x))
yfit<-yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue")
```



Q-Q Plot

```
qqnorm(fieldanovares,main="Normal Q-Q Plot, field anova residuals")
qqline(fieldanovares,col="blue")
```



Normality assumption looks ok. The stepwise appearance of the probability plot is due to the ordinal data. Shapiro test does not work due to the sample size being greater than 5000.

Follow-up on ANOVA results:

TukeyHSD Pairwise Test:

TukeyHSD pairwise test:
We used alpha = 0.05 for this analysis.

> **H_0**: No relationship between fields in terms of "like" rating
> **H_a**: Relationship between fields in terms of "like" rating

P-values ≤0.05 support Ha – there is a statistically significant relationship. "Diff" column shows the difference in like rating between each pair of academic fields.

```
TukeyHSD(fieldanova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = like_o ~ field_cd)
##
## $field_cd
##                 diff          lwr         upr     p adj
## 2-1   -5.477380e-01 -1.062013020 -0.03346299 0.0231535
## 3-1   -1.036165e-01 -0.455096743  0.24786368 0.9999179
## 4-1    6.054404e-01  0.007063375  1.20381734 0.0437274
## 5-1   -5.032553e-01 -0.837814792 -0.16869580 0.0000233
## 6-1   -1.225576e-01 -0.559913461  0.31479820 0.9999597
## 7-1   -3.388494e-02 -0.517577399  0.44980752 1.0000000
## 8-1    7.162393e-02 -0.219207359  0.36245522 0.9999938
```

```
## 9-1   -1.184489e-01 -0.478340494  0.24144264 0.9996353
## 10-1  -1.079943e-01 -0.431073602  0.21508493 0.9995530
## 11-1  -1.226397e-01 -0.511378538  0.26609905 0.9997925
## 12-1  -4.825911e-02 -1.534152229  1.43763401 1.0000000
## 13-1   1.283890e-01 -0.220696676  0.47747462 0.9984937
## 14-1  -9.549811e-02 -0.725410350  0.53441413 1.0000000
## 15-1  -2.473303e-01 -0.797293705  0.30263307 0.9850532
## 16-1   6.688462e-01 -0.371183837  1.70887615 0.7235449
## 17-1   4.938462e-01 -1.540492641  2.52818495 0.9999950
## 18-1  -1.245212e-01 -1.070304972  0.82126259 1.0000000
## 3-2    4.441215e-01 -0.068330476  0.95657342 0.1885076
## 4-2    1.153178e+00  0.448108384  1.85824834 0.0000018
## 5-2    4.448271e-02 -0.456515013  0.54548043 1.0000000
## 6-2    4.251804e-01 -0.149579387  0.99994013 0.4674511
## 7-2    5.138531e-01 -0.096906344  1.12461248 0.2326663
## 8-2    6.193619e-01  0.146443867  1.09228001 0.0006944
## 9-2    4.292891e-01 -0.088968195  0.94754635 0.2577933
## 10-2   4.397437e-01 -0.053661715  0.93314905 0.1516535
## 11-2   4.250983e-01 -0.113591363  0.96378788 0.3431284
## 12-2   4.994789e-01 -1.032495982  2.03145377 0.9996780
## 13-2   6.761270e-01  0.165314437  1.18693951 0.0005535
## 14-2   4.522399e-01 -0.279783391  1.18426318 0.7826289
## 15-2   3.004077e-01 -0.364068338  0.96488371 0.9842142
## 16-2   1.216584e+00  0.111717179  2.32145114 0.0146459
## 17-2   1.041584e+00 -1.026652529  3.10982085 0.9549775
## 18-2   4.232168e-01 -0.593432399  1.43986602 0.9935142
## 4-3    7.090569e-01  0.112246015  1.30586776 0.0044484
## 5-3   -3.996388e-01 -0.731389066 -0.06788846 0.0034984
## 6-3   -1.894110e-02 -0.454151770  0.41626957 1.0000000
## 7-3    6.973159e-02 -0.412022081  0.55148527 1.0000000
## 8-3    1.752405e-01 -0.112354818  0.46283575 0.8009409
## 9-3   -1.483240e-02 -0.372114001  0.34244921 1.0000000
## 10-3  -4.377804e-03 -0.324547163  0.31579155 1.0000000
## 11-3  -1.902321e-02 -0.405346983  0.36730056 1.0000000
## 12-3   5.535742e-02 -1.429905709  1.54062056 1.0000000
## 13-3   2.320055e-01 -0.114388775  0.57839978 0.6557004
## 14-3   8.118424e-03 -0.620306304  0.63654315 1.0000000
## 15-3  -1.437138e-01 -0.691972788  0.40454522 0.9999845
## 16-3   7.724627e-01 -0.266667040  1.81159241 0.4578716
## 17-3   5.974627e-01 -1.436416007  2.63134138 0.9999219
## 18-3  -2.090466e-02 -0.965698375  0.92388905 1.0000000
## 5-4   -1.108696e+00 -1.695700710 -0.52169059 0.0000000
## 6-4   -7.279980e-01 -1.379093222 -0.07690276 0.0117129
## 7-4   -6.393253e-01 -1.322408949  0.04375836 0.0988718
## 8-4   -5.338164e-01 -1.097046148  0.02941330 0.0876888
## 9-4   -7.238893e-01 -1.325692259 -0.12208631 0.0035901
## 10-4  -7.134347e-01 -1.293973317 -0.13289607 0.0024417
## 11-4  -7.280801e-01 -1.347565962 -0.10859424 0.0053391
## 12-4  -6.536995e-01 -2.215916879  0.90851795 0.9931284
## 13-4  -4.770514e-01 -1.072455174  0.11835240 0.3153484
## 14-4  -7.009385e-01 -1.494307429  0.09243050 0.1626387
## 15-4  -8.527707e-01 -1.584279122 -0.12126223 0.0061074
## 16-4   6.340580e-02 -1.083026610  1.20983820 1.0000000
## 17-4  -1.115942e-01 -2.202330727  1.97914232 1.0000000
```

```
## 18-4  -7.299615e-01 -1.791635623  0.33171252 0.6083758
## 6-5    3.806977e-01 -0.040965765  0.80236109 0.1362474
## 7-5    4.693704e-01 -0.000180835  0.93892155 0.0502170
## 8-5    5.748792e-01  0.308228210  0.84153024 0.0000000
## 9-5    3.848064e-01  0.044157186  0.72545555 0.0100410
## 10-5   3.952610e-01  0.093764463  0.69675745 0.0006796
## 11-5   3.806155e-01  0.009619897  0.75161120 0.0371232
## 12-5   4.549962e-01 -1.026353972  1.93634634 0.9998556
## 13-5   6.316443e-01  0.302432003  0.96085653 0.0000000
## 14-5   4.077572e-01 -0.211362638  1.02687701 0.6846641
## 15-5   2.559250e-01 -0.281643300  0.79349326 0.9734699
## 16-5   1.172101e+00  0.138572394  2.20563050 0.0094553
## 17-5   9.971014e-01 -1.033921509  3.02812441 0.9644075
## 18-5   3.787341e-01 -0.559896220  1.31736442 0.9954496
## 7-6    8.867269e-02 -0.458893181  0.63623857 1.0000000
## 8-6    1.941816e-01 -0.193700931  0.58206406 0.9573524
## 9-6    4.108704e-03 -0.437922889  0.44614030 1.0000000
## 10-6   1.456329e-02 -0.398050570  0.42717716 1.0000000
## 11-6  -8.211473e-05 -0.465901649  0.46573742 1.0000000
## 12-6   7.429852e-02 -1.433596817  1.58219386 1.0000000
## 13-6   2.509466e-01 -0.182332498  0.68422570 0.8593231
## 14-6   2.705952e-02 -0.653131202  0.70725025 1.0000000
## 15-6  -1.247727e-01 -0.731674422  0.48212905 0.9999996
## 16-6   7.914038e-01 -0.279825589  1.86263316 0.4700319
## 17-6   6.164038e-01 -1.434060631  2.66686820 0.9998925
## 18-6  -1.963561e-03 -0.981951800  0.97802468 1.0000000
## 8-7    1.055089e-01 -0.333957971  0.54497571 0.9999957
## 9-7   -8.456399e-02 -0.572488373  0.40336039 1.0000000
## 10-7  -7.410940e-02 -0.535551135  0.38733234 1.0000000
## 11-7  -8.875481e-02 -0.598329251  0.42081964 1.0000000
## 12-7  -1.437417e-02 -1.536355323  1.50760698 1.0000000
## 13-7   1.622739e-01 -0.317735522  0.64228334 0.9994821
## 14-7  -6.161317e-02 -0.772484253  0.64925791 1.0000000
## 15-7  -2.134454e-01 -0.854544429  0.42765367 0.9995757
## 16-7   7.027311e-01 -0.388236665  1.79369885 0.7210113
## 17-7   5.277311e-01 -1.533114025  2.58857621 0.9999890
## 18-7  -9.063625e-02 -1.092162700  0.91089019 1.0000000
## 9-8   -1.900729e-01 -0.487889325  0.10774361 0.7354001
## 10-8  -1.796183e-01 -0.431715313  0.07247878 0.5407982
## 11-8  -1.942637e-01 -0.526363898  0.13783654 0.8490112
## 12-8  -1.198830e-01 -1.591973756  1.35220767 1.0000000
## 13-8   5.676504e-02 -0.227898792  0.34142887 0.9999997
## 14-8  -1.671220e-01 -0.763747668  0.42950359 0.9999599
## 15-8  -3.189542e-01 -0.830454431  0.19254593 0.7697012
## 16-8   5.972222e-01 -0.422991035  1.61743548 0.8482179
## 17-8   4.222222e-01 -1.602057166  2.44650161 0.9999995
## 18-8  -1.961451e-01 -1.120092994  0.72780274 0.9999993
## 10-9   1.045459e-02 -0.318926676  0.33983586 1.0000000
## 11-9  -4.190818e-03 -0.398182767  0.38980113 1.0000000
## 12-9   7.018982e-02 -1.417086273  1.55746591 1.0000000
## 13-9   2.468379e-01 -0.108088286  0.60176408 0.5870322
## 14-9   2.295082e-02 -0.610216804  0.65611844 1.0000000
## 15-9  -1.288814e-01 -0.682570413  0.42480764 0.9999973
## 16-9   7.872951e-01 -0.254709808  1.82929997 0.4260808
```

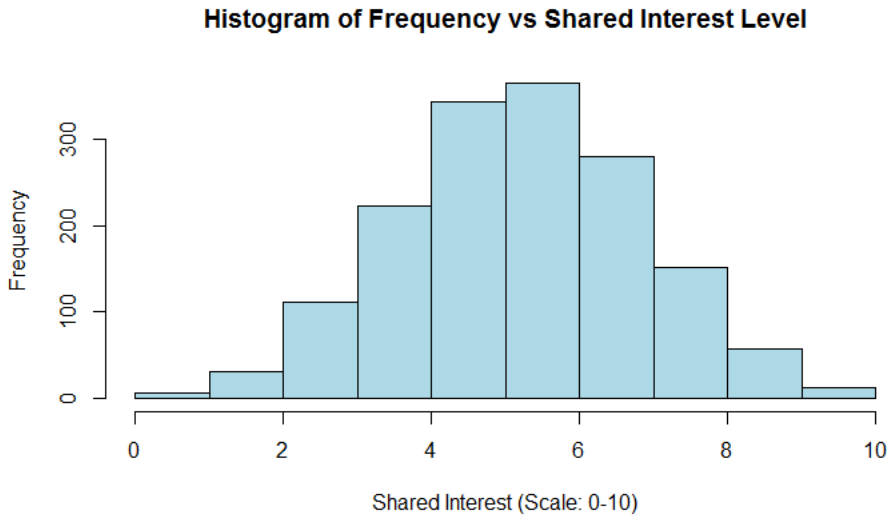
```
## 17-9   6.122951e-01 -1.423054063  2.64764423 0.9998914
## 18-9  -6.072265e-03 -0.954027309  0.94188278 1.0000000
## 11-10 -1.464541e-02 -0.375322451  0.34603163 1.0000000
## 12-10  5.973523e-02 -1.419064436  1.53853489 1.0000000
## 13-10  2.363833e-01 -0.081155460  0.55392207 0.4551074
## 14-10  1.249623e-02 -0.600496033  0.62548849 1.0000000
## 15-10 -1.393360e-01 -0.669835564  0.39116360 0.9999841
## 16-10  7.768405e-01 -0.253029629  1.80671061 0.4293050
## 17-10  6.018405e-01 -1.427322986  2.63100397 0.9999107
## 18-10 -1.652686e-02 -0.951126790  0.91807308 1.0000000
## 12-11  7.438064e-02 -1.420138042  1.56889932 1.0000000
## 13-11  2.510287e-01 -0.133117747  0.63517518 0.6977349
## 14-11  2.714164e-02 -0.622856169  0.67713944 1.0000000
## 15-11 -1.246906e-01 -0.697549580  0.44816844 0.9999990
## 16-11  7.914859e-01 -0.260830639  1.84380244 0.4349338
## 17-11  6.164859e-01 -1.424161558  2.65713336 0.9998849
## 18-11 -1.881447e-03 -0.961159651  0.95739676 1.0000000
## 13-12  1.766481e-01 -1.308050214  1.66134637 1.0000000
## 14-12 -4.723900e-02 -1.621804865  1.52732687 1.0000000
## 15-12 -1.990712e-01 -1.743392405  1.34524999 1.0000000
## 16-12  7.171053e-01 -1.061695067  2.49590559 0.9954957
## 17-12  5.421053e-01 -1.952088508  3.03629903 0.9999991
## 18-12 -7.626208e-02 -1.801653399  1.64912923 1.0000000
## 14-13 -2.238871e-01 -0.850975659  0.40320150 0.9989594
## 15-13 -3.757193e-01 -0.922446261  0.17100769 0.6092948
## 16-13  5.404572e-01 -0.497865039  1.57877941 0.9396957
## 17-13  3.654572e-01 -1.668009066  2.39892343 1.0000000
## 18-13 -2.529102e-01 -1.196815674  0.69099535 0.9999788
## 15-14 -1.518322e-01 -0.909353844  0.60568943 0.9999997
## 16-14  7.643443e-01 -0.398858937  1.92754746 0.6885117
## 17-14  5.893443e-01 -1.510635174  2.68932370 0.9999588
## 18-14 -2.902308e-02 -1.108785215  1.05073905 1.0000000
## 16-15  9.161765e-01 -0.205746869  2.03809981 0.2818384
## 17-15  7.411765e-01 -1.336221886  2.81857483 0.9989683
## 18-15  1.228091e-01 -0.912351035  1.15796928 1.0000000
## 17-16 -1.750000e-01 -2.432163393  2.08216339 1.0000000
## 18-16 -7.933673e-01 -2.153793460  0.56705877 0.8522337
## 18-17 -6.183673e-01 -2.833684742  1.59695005 0.9999619
```

ii. *Do opposites attract or birds of a feather flock together?*

The Mean Like Rating for ‘Low Shared Interest Couples’ was compared against ‘High Shared Interest Couples’ through Two Sample t-Test. We calculated the t-Statistic for the two groups and found a statistically significant difference.

```
attach(date)
date['mean_shar'] <- (date['shar'] + date['shar_o'])/2
```

```
uniqueIndex <- !duplicated(date['IidPid'])

unique <- date[uniqueIndex,]

subset <- unique[!is.na(unique$mean_shar),]
summary(subset$mean_shar)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   4.500   5.500   5.449   6.500  10.000

hist(subset$mean_shar, col='light blue', xlab='Shared Interest (Scale: 0-1
0)', ylab='Frequency', main='Histogram of Frequency vs Shared Interest Lev
el')
```




```
subset2 <- date[!is.na(date$mean_shar),]
lowSharedInt <- date[date$mean_shar < 5.5,]
highSharedInt <- date[date$mean_shar >= 5.5,]
```

Summary Statistics of Overall Rating received by date in Low Shared Interest Group

```
summary(lowSharedInt$like_o)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   4.000   5.000   5.277   7.000  10.000    1895
```

Summary Statistics of Overall Rating received by date in High Shared Interest Group

```
summary(highSharedInt$like_o)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   7.000   6.836   8.000  10.000    1900
```

Two Sample t-Test

H_0: $\mu_{\text{Low Shared Interest Couples}} = \mu_{\text{High Shared Interest Couples}}$

H_a: $\mu_{\text{Low Shared Interest Couples}} \neq \mu_{\text{High Shared Interest Couples}}$

```
t.test(lowSharedInt$like_o, highSharedInt$like_o)

##
##  Welch Two Sample t-test
##
## data:  lowSharedInt$like_o and highSharedInt$like_o
## t = -36.259, df = 5433.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.643792 -1.475160
## sample estimates:
## mean of x mean of y
##  5.276865  6.836341
```

Two Sample t-Test Result

The p-Value for the test came out to be very low i.e. less than 2.2×10^{-6}. Thus, it is extremely unlikely that the two samples came from the same population.

We will reject H_0 and accept H_a. Hence, it is safe to infer that Mean Like Rating among 'Low Shared Interest Couples' is different from 'High Shared Interest Couples'. In the latter group subjects have received statistically greater Mean Like Rating from the partners.

Assumptions for Two Sample t-Test

- **Random Sampling** – We will have to assume that the participants in the Speed Dating Experiment are randomly sampled
- **Stability** – The characteristics of the process of Speed Dating between the Subject and Partner remain stable over time
- **Normality** – Since, the number of observations are greater than 30 for each of the two groups Central Limit Theorem applies and the assumption for normality is not required

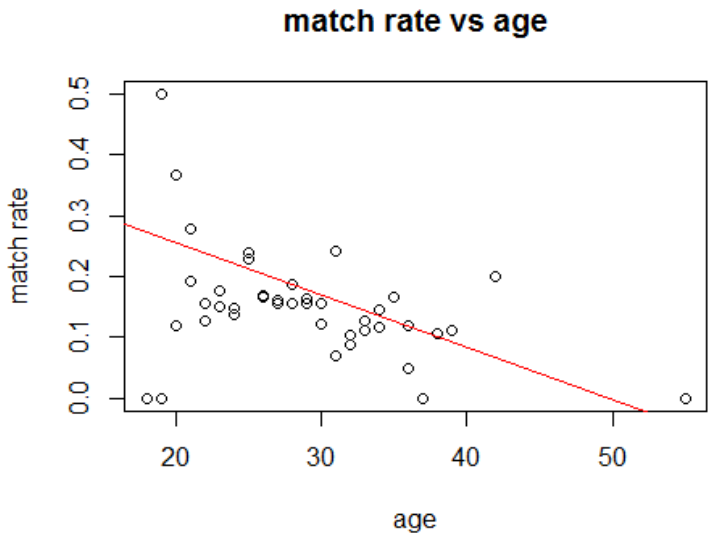
iii. *How is age associated with the possibility of finding a partner for males? For females?*

Multiple Regression

H_0: No relationship of Match Rate with Age*Gender

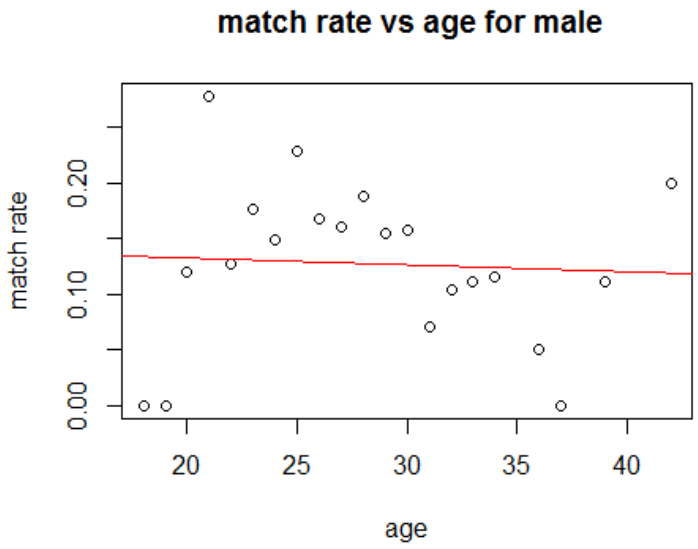
H_a: Relationship

```
r_match <- aggregate(match~age+gender,date,mean)
r1 <- lm(r_match$match ~ r_match$age*r_match$gender)
summary(r1)

##
## Call:
## lm(formula = r_match$match ~ r_match$age * r_match$gender)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.133323 -0.046637 -0.007715  0.038619  0.235945
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.426756   0.065744   6.491 1.36e-07 ***
## r_match$age                 -0.008563   0.002158  -3.968  0.00032 ***
## r_match$gender              -0.282560   0.098669  -2.864  0.00686 **
## r_match$age:r_match$gender   0.007959   0.003320   2.397  0.02169 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07716 on 37 degrees of freedom
## Multiple R-squared:  0.3481, Adjusted R-squared:  0.2953
## F-statistic: 6.587 on 3 and 37 DF,  p-value: 0.001114

plot(r_match$age,r_match$match,xlab = 'age',ylab = 'match rate',main = 'ma
tch rate vs age')
abline(r1,lyt=2,col='red')

## Warning in abline(r1, lyt = 2, col = "red"): only using the first two o
f 4
## regression coefficients

## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): "l
yt"
## is not a graphical parameter
```



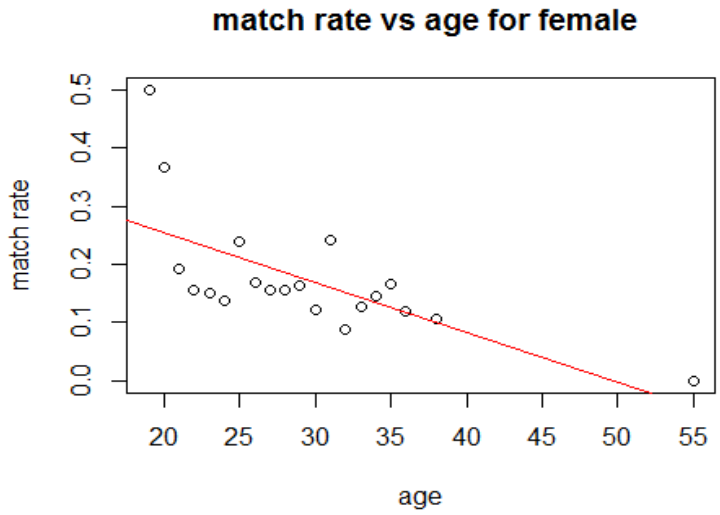
Match rate with age for male

```
sub1 <- subset(r_match,gender==1)
r2 <- lm(sub1$match~sub1$age)
summary(r2)

##
## Call:
## lm(formula = sub1$match ~ sub1$age)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.133323 -0.020290 -0.003759  0.038619  0.146267
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1441962  0.0717459   2.010   0.0589 .
## sub1$age    -0.0006041  0.0024609  -0.245   0.8087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07525 on 19 degrees of freedom
## Multiple R-squared:  0.003161,   Adjusted R-squared:  -0.0493
## F-statistic: 0.06026 on 1 and 19 DF,  p-value: 0.8087

plot(sub1$age,sub1$match,xlab = 'age',ylab = 'match rate',main = ' match r
ate vs age for male')
abline(r2,col='red')
```



Match rate with age for female

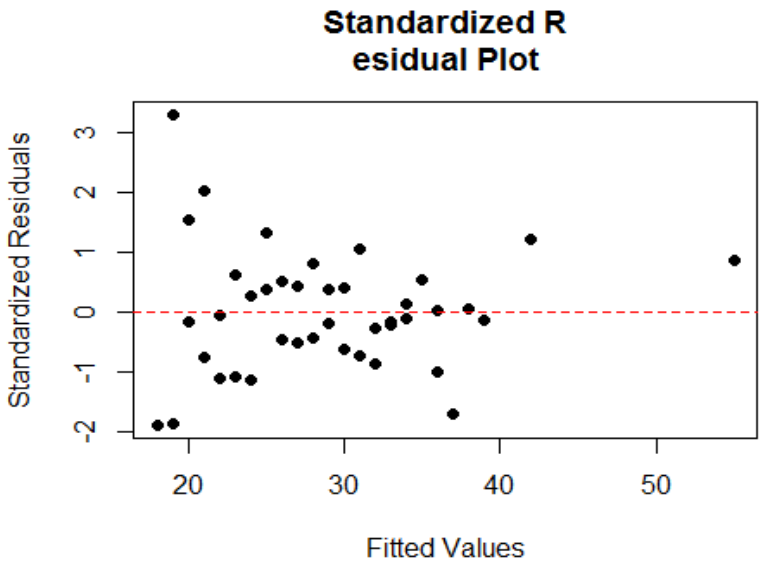
```
sub2 <- subset(r_match,gender==0)
r3 <- lm(sub2$match~sub2$age)
summary(r3)

##
## Call:
## lm(formula = sub2$match ~ sub2$age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08468 -0.04890 -0.01580  0.03007  0.23594
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.426756   0.067426   6.329 5.78e-06 ***
## sub2$age    -0.008563   0.002213  -3.869  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07914 on 18 degrees of freedom
## Multiple R-squared:  0.4541, Adjusted R-squared:  0.4238
## F-statistic: 14.97 on 1 and 18 DF,  p-value: 0.001123

plot(sub2$age,sub2$match,xlab = 'age',ylab = 'match rate',main = ' match r
ate vs age for female')
abline(r3,col='red')
```



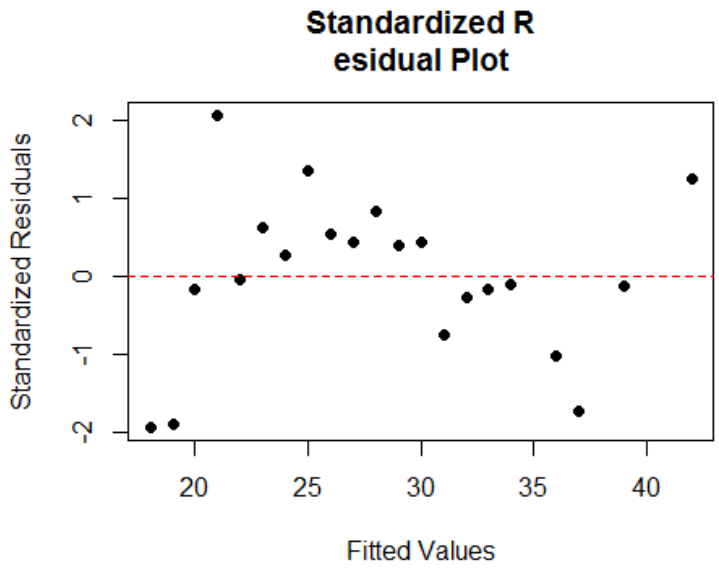
Assumptions for Regression

- **Random Sampling** – We will have to assume that the participants in the Speed Dating Experiment are randomly sampled
- **Stability** – The characteristics of the process of Speed Dating between the Subject and Partner remain stable over time
- **Standard Residuals ~ Normal (Mean 0, SD σ)** *constant*
  Over all possible values of the predictors

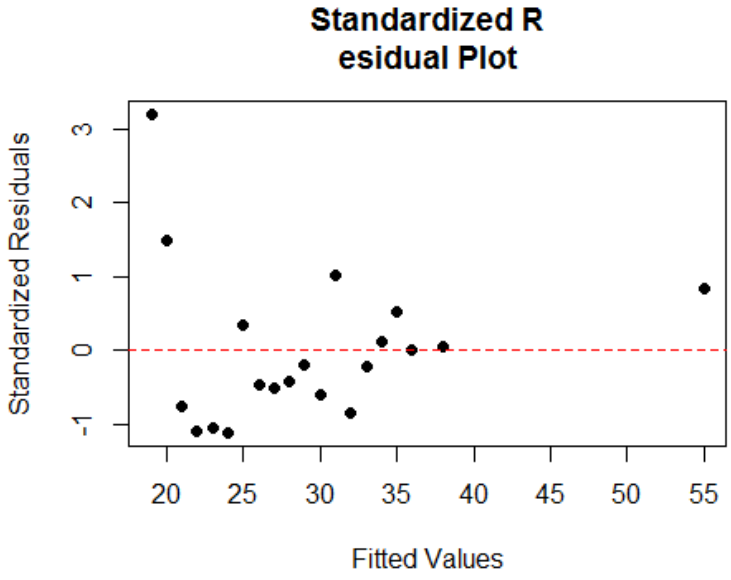
```
r1.stres<-rstandard(r1)
plot(r_match$age,r1.stres,pch = 16, main = "Standardized R
esidual Plot", xlab = "Fitted Values", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="red")
```



```
r2.stres<-rstandard(r2)
plot(sub1$age,r2.stres,pch = 16, main = "Standardized R
esidual Plot", xlab = "Fitted Values", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="red")
```

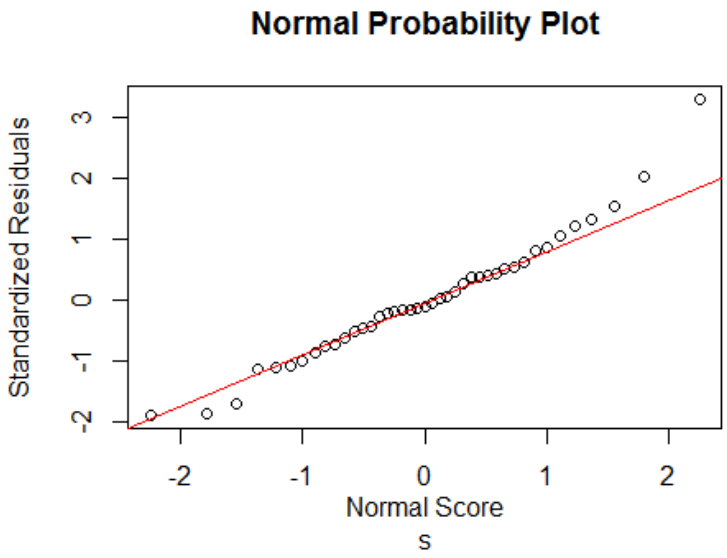



```
r3.stres<-rstandard(r3)
plot(sub2$age,r3.stres,pch = 16, main = "Standardized R
esidual Plot", xlab = "Fitted Values", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="red")
```

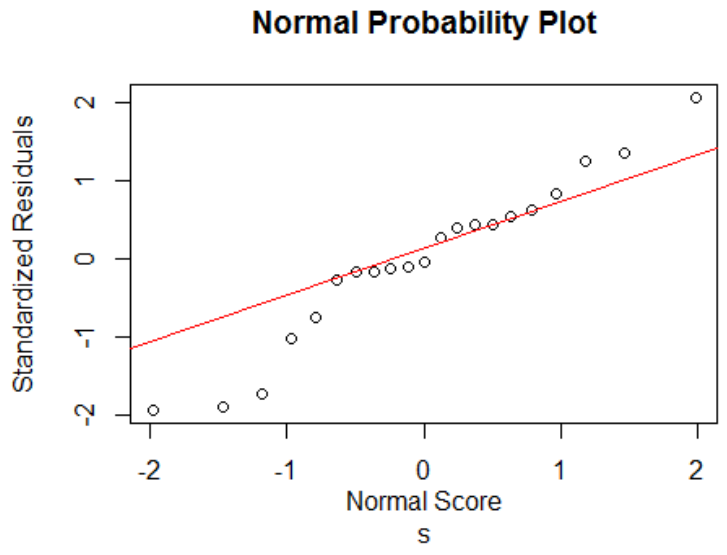


Plots

```
qqnorm(r1.stres, main = "Normal Probability Plot", xlab = "Normal Score
s", ylab = "Standardized Residuals")
qqline(r1.stres, col = "red")
```

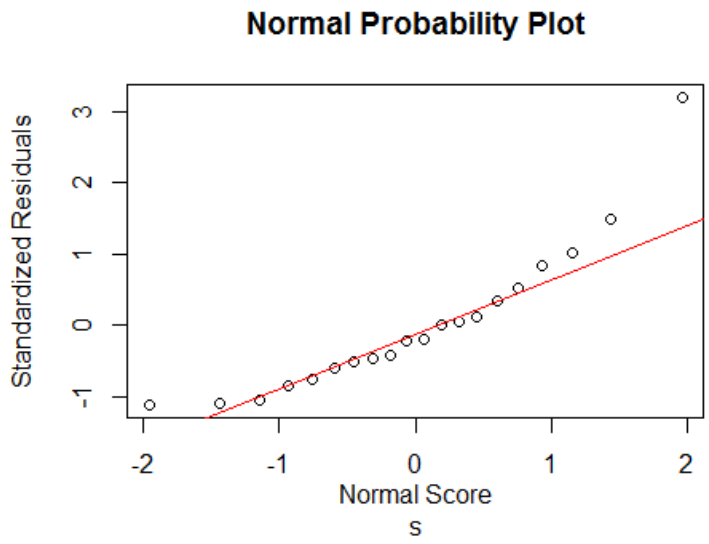



```
shapiro.test(r1.stres)
##  Shapiro-Wilk normality test
##
## data:  r1.stres
## W = 0.9656, p-value = 0.2458

qqnorm(r2.stres, main = "Normal Probability Plot", xlab = "Normal Score
s", ylab = "Standardized Residuals")
qqline(r2.stres, col = "red")
```




```
shapiro.test(r2.stres)
##  Shapiro-Wilk normality test
##
## data:  r2.stres
## W = 0.95587, p-value = 0.4372
```

```
qqnorm(r3.stres, main = "Normal Probability Plot", xlab = "Normal Score
s", ylab = "Standardized Residuals")
qqline(r3.stres, col = "red")
```




```
shapiro.test(r3.stres)
##  Shapiro-Wilk normality test
##
## data:  r3.stres
## W = 0.86488, p-value = 0.009561
```

iv. *What's the gender difference in mate selection? – Attributes Difference*

Multiple Regression

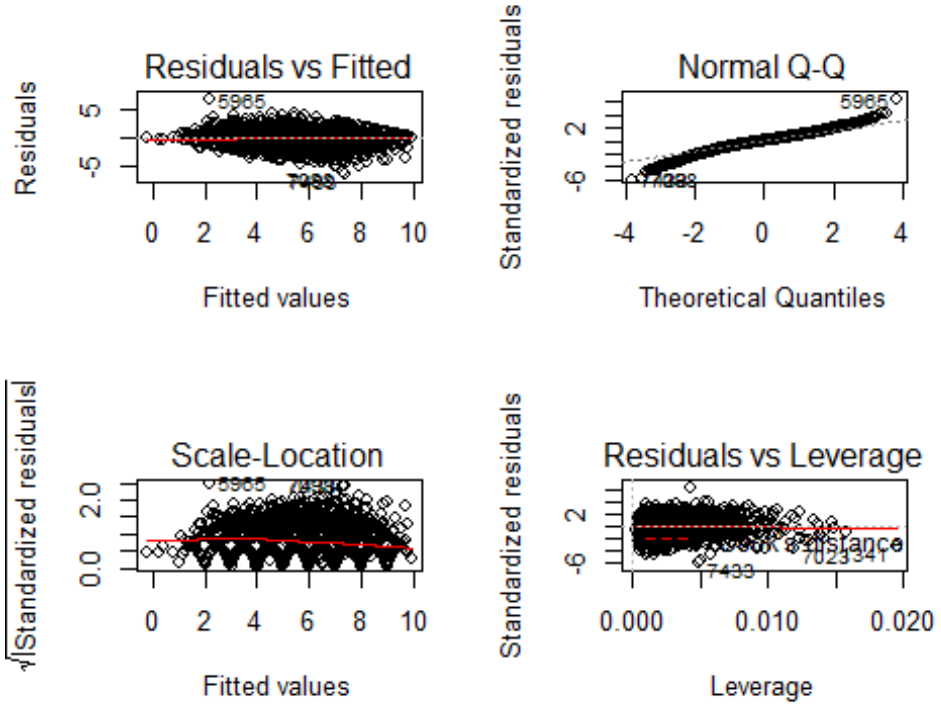
We checked the relationship of 'Like Rating Received' with the Interaction of Gender with different Attributes

```
fit2 = lm(like_o ~ gender * attr_o + gender * sinc_o + gender * intel_o +
gender * fun_o + gender * amb_o + gender * shar_o )
summary(fit2)

##
## Call:
## lm(formula = like_o ~ gender * attr_o + gender * sinc_o + gender *
##     intel_o + gender * fun_o + gender * amb_o + gender * shar_o)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -6.3805 -0.5780  0.0768  0.6350  6.8048
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.430e-01  9.834e-02   2.471 0.013513 *
## gender         -4.628e-01  1.345e-01  -3.440 0.000585 ***
## attr_o          3.428e-01  1.219e-02  28.129  < 2e-16 ***
## sinc_o          8.988e-02  1.549e-02   5.803 6.82e-09 ***
## intel_o         7.844e-02  1.830e-02   4.287 1.84e-05 ***
## fun_o           2.155e-01  1.436e-02  15.001  < 2e-16 ***
## amb_o          -3.285e-02  1.413e-02  -2.325 0.020103 *
## shar_o          2.525e-01  1.113e-02  22.693  < 2e-16 ***
## gender:attr_o  -8.173e-02  1.702e-02  -4.802 1.60e-06 ***
## gender:sinc_o   3.582e-02  2.089e-02   1.715 0.086407 .
## gender:intel_o  5.581e-02  2.558e-02   2.181 0.029185 *
## gender:fun_o    1.287e-02  1.976e-02   0.652 0.514686
## gender:amb_o    3.077e-02  1.955e-02   1.574 0.115615
## gender:shar_o  -7.742e-05  1.582e-02  -0.005 0.996096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.078 on 6994 degrees of freedom
##   (1370 observations deleted due to missingness)
## Multiple R-squared:  0.6617, Adjusted R-squared:  0.661
## F-statistic:  1052 on 13 and 6994 DF,  p-value: < 2.2e-16
```

Since, we are getting significant p-Values for **gender:attr_o** and **gender:intel:o,** we can conclude that there is a relationship of these interaction variables with 'Like Rating Received'. Hence, Attractiveness and Intelligence of subjects drive their overall Like Rating Received by their partners to a different extent depending on subject's gender.

```
par(mfrow = c(2,2))
plot(fit2)
```



Looks Good

v. *Are women more selective then men?*

```
t.test(m,f)

##
##  Welch Two Sample t-test
##
## data:  m and f
## t = -6.9772, df = 8062.9, p-value = 3.249e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3641070 -0.2043875
## sample estimates:
## mean of x mean of y
##  5.991850  6.276097
```

vi. *What influences love at first sight?*

Finally, we have tried to arrive at the best model using all significant factors trying to maximize R^2 while minimizing S and at the same time aiming for simplest model which captures most variation in Like Ratings Received.

Multiple Regression

H_0: No relationship

H_a: Relationship

The p-Values for all predictors is significant and all of the $\hat{\beta} \neq 0$. We can safely reject **H_0**.

Our model captures 66.13% of the variation in 'Like Rating Received' by an individual in this sample.

```
fit3 = lm(data = date, like_o ~  gender :attr_o  + gender :intel_o  + attr
_o + sinc_o + intel_o + fun_o  + shar_o +age:gender)


summary(fit3)

##
## Call:
## lm(formula = like_o ~ gender:attr_o + gender:intel_o + attr_o +
##     sinc_o + intel_o + fun_o + shar_o + age:gender, data = date)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3548 -0.5795  0.0741  0.6464  6.8540
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.141901   0.083286   1.704 0.088465 .
## attr_o          0.331488   0.011236  29.503  < 2e-16 ***
## sinc_o          0.112172   0.010306  10.884  < 2e-16 ***
## intel_o         0.049560   0.014323   3.460 0.000543 ***
## fun_o           0.220462   0.009655  22.835  < 2e-16 ***
## shar_o          0.248379   0.007763  31.996  < 2e-16 ***
## gender:attr_o  -0.063814   0.014200  -4.494 7.10e-06 ***
## gender:intel_o  0.095694   0.015691   6.099 1.13e-09 ***
## gender:age     -0.012379   0.003604  -3.434 0.000597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.078 on 7091 degrees of freedom
##   (1278 observations deleted due to missingness)
```

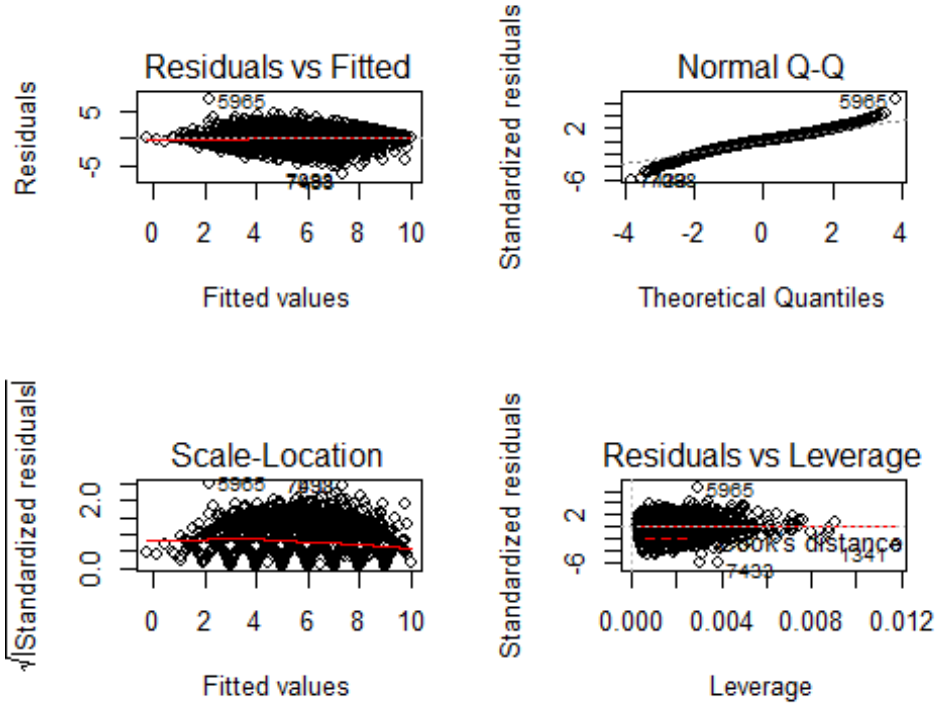
```
## Multiple R-squared:  0.6613, Adjusted R-squared:  0.661
## F-statistic:  1731 on 8 and 7091 DF,  p-value: < 2.2e-16
```

Assumptions for Regression

- **Random Sampling** – We will have to assume that the participants in the Speed Dating Experiment are randomly sampled
- **Stability** – The characteristics of the process of Speed Dating between the Subject and Partner remain stable over time
- **Standard Residuals ~ Normal (Mean 0, SD σ)** *constant*
  Over all possible values of the predictors

Assumption Check

```
par(mfrow = c(2,2))
plot(fit3)
```



Looks good in general, but have some problems in the tails

Multiple Regression is a robust statistical technique that allows us to show good tolerance with assumption validity. So, our model still holds good.

7. REFERENCES

Data set:

http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/

Accessed July & August 2016

Compiled by Ray Fisman and Sheena Iyengar, Columbia Business School

Final paper: Fisman, Raymond; Iyengar, Sheena S.; Kamenica, Emir; Simonson, Itamar. "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment"

The Quarterly Journal of Economics, 1 May 2006, Vol.121(2), pp.673-697

Treating 0-10 "like" rating (ordinal data) as continuous:

Menard (2002): "When the dependent variable is ordinal but has five or more categories, treating the variable as though it were continuous is acceptable."

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.