unlocking the secrets behind adult dating

DECEMBER 19, 2021

YASH SOMAIYA & LYNN AHRENS

GITHUB REPOSITORY

prepared for the shareholders @ bumble

TABLE OF CONTENTS

THE BASICS

DATA

MODELING

CONCLUSIONS

APPENDIX

EXECUTIVE SUMMARY

In the modern dating world, combined with the rapid onset of technology, adults are now faced with both a plethora of decisions and a constant influx of information about potential matches. In this quest for love, with so many factors to consider, this can make an already daunting position more intimidating and overwhelming. In the following analyses, we seek to take data from over 8,000 speed dating pairings to form a predictive model for match rate based on selected human characteristics, preferences, goals, and reception to others.

Our data is taken from an original experimental study conducted in Columbia University by Raymond Fisman, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson, where students of the graduate and professional schools attended speed dating events in the area between the years 2002 and 2004. Recorded are data for each pairing, ranging from field of study and hobbies to frequency of dates and ratings of their partners that round. Our explanatory variables span four core divisions: demographic information (e.g. race, gender), hobbies and activities (e.g. hiking, shopping), preferences, and external feedback (e.g. partner's ratings), all used in prediction of the final response variable, or the 'yes' frequency rate of an individual in the dating pool.

Our data was subsequently split into a training set for model fitting and a test set for model assessment, and our data exploratory phase consists of determination of variation within the response variable/features, and initial insights into the covariation between features and a relationship between the features and response variable. In finalizing our predictive model, we built five different cross-validated models: ordinary least squares, ridge regression, LASSO regression, random forest and a pruned decision tree. Of all of the regression models, and overall, the OLS model had the lowest test error.

The strongest predictor of match percentage was found to be attractiveness, exceeding the next predictor, gender, by a significant amount, and followed by sincerity, fun, shared interest ratings. Within the context of the speed dating experiment, this was supported by our intuition and best judgement given this is the most easily recognizable trait and can be widely interpreted for each participant. For our stakeholders at Bumble, we advise greater emphasis on recommendations by shared interests, verifying metrics for attraction by user, and continuing with its policy of women initiating conversation for more successful odds, given that women are found to be more selective by our model than men. It is our goal with this analysis to cultivate effortless, enjoyable, and healthy matches in today's dating climate.

INTRODUCTION

In the United States alone, there are over 124.6 million American adults who are single, approximately 50.2% of the entire population. With the onset of technology and digital connections, it has become far easier for singles to formalize desired characteristics, gain awareness of potential options, screen matches, and conceptualize a lifelong partner prior to a dating phase.[1] According to Eric Klinenberg, a New York University sociology professor, 'one big issue is people today are really looking for their soulmate, and they're not going to compromise.' This onset of selection and options lends itself to increasingly selective dating pools and single participants. However, according to a 2019 Pew Research Study conducted prior to the pandemic, single Americans have been found to be generally open to dating those with different backgrounds from their own, including those of different races, ethnicities, and religions, though were united in identifying concrete dealbreakers for creating a match, including large geographic displacement, >10 year age gap, and with few shared interests or willingness to try new things .[1]

Interestingly, with an extreme culture change around dating moving into the 20th century, adults are more inclined to stay single and push off marriage in favor of investing valuable time into their careers and establishing a productive solitary routine in impressionable early years.[2] Ultimately, this contemporary culture has lent itself to exploration of new people, experiences, and compatibility potential.[3]

Given our understanding of a variety of factors to affect compatibility, preference, and selection within the current dating pool in adults, we sought to investigate the extent to which each selected characteristic, and often the degree of difference between the two individuals, play a role in the final match outcome within a speed dating environment, ranging from race/ethnicity to goals in dating, future career, mean income of hometown zipcode, frequency of social activities, etc, all of which will be used to predict the compatibility of a pairing of single individuals, the response being a match or no match. The success of our analysis will be measured in successful prediction of match outcome data by >50% in the test data pairings. It is our hope that our analysis may provide a predictive model of compatibility based on a vast extraction of priorities, ideals, and traits, and contribute to the growing body of research on single adult life dynamics, interactions, and motivations in our modern day lives. Our research may also shed light on how favorability and love is affected by difference in background or motive, and ultimately, expand upon current considerations of predetermined associations or preferences for selected traits.

1 Brown, Susan L., and Sayaka K. Shinohara. "Dating Relationships in Older Adulthood: A National Portrait." Journal of Marriage and Family 75, no. 5 (2013): 1194–1202. http://www.jstor.org/stable/24583366

2 Brown, Anna. "Americans' Views on Dating and Relationships." Pew Research Center's Social & Demographic Trends Project, Pew Research Center, 2 Oct. 2020, https://www.pewresearch.org/social-trends/2020/08/20/nearly-half-of-u-s-adults-say-dating-has-gotten-harder-for-most-people-in-the-last-10-years/

3 Xia, Mengya et al. "A Developmental Perspective on Young Adult Romantic Relationships: Examining Family and Individual Factors in Adolescence." Journal of youth and adolescence vol. 47,7 (2018): 1499-1516. doi:10.1007/s10964-018-0815-8

DATA: SOURCE

The data were collected within the context of a Speed Dating experiment conducted by Raymond Fisman, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson, where each pairing of participants was asked to engage in a four-minute conversation to determine compatibility and sociability for themselves and then given one minute to fill out their scorecards. If both partners were interested in meeting each other a second time, they gave an 'accept' response, and contact information was exchanged between both parties.

The subjects were graduate and professional school students at Columbia University, recruited through fliers and emails distributed throughout campus and through research assistant connections. In New York alone, over eight companies are devoted to hosting dating events in the speed dating service, and this premise was used to inspire the setup of the speed dating experiment. Each event hosted within this study had the same setup with regard to table arrangement, lighting, and type/volume of music. Between 2002-2004, these speed dating events were hosted on weekday evenings, participants were randomly distributed and divided, and not made aware of the number of new participants they would be meeting that night.

Each subject was given an ID number and a clipboard on which to record their response of 'yes' or 'no' with respect to wanting to meet their partner again, whose ID they were requested to denote on their form, in addition to rate questions about their partner's attributes (attractiveness, sincerity, intelligence, fun, ambition, and shared interest). This data was recorded and subsequently combined with each participant's pre and post event surveys, the former which inquired of basic demographic information, and the latter asking for general event experience feedback. The data online can be found here.

DATA: ALLOCATION

Our data was divided utilizing an 80-20 split, indicating that 80% of our data was allocated to a training dataset, while the remaining 20% was denoted as the test set. A random seed was put in place to ensure consistency between runs.

DATA: CLEANING PROCESS

Our data cleaning process primarily consisted of removing unneeded features from our dataset. The original dataset consisted of 195 features, many of which were not applicable to our analysis or contained too many incomplete values. Once these unwanted features and empty data points were eliminated, though, we still had thorough data massaging to perform.

The original dataset presented the data where each observation corresponded to each individual report a subject filled out on one of their partners. However, to analyze individual subjects' attributes effect on their frequency of yeses from partners, each report contained the evaluatee's individual ID number, so we compiled and averaged the ratings given to each individual subject by their partners. We were then able to eliminate duplicate rows.

DATA: DESCRIPTION

Observations

In the pre-cleaning state, the data are structured by unique pairings. There are 8,378 observations, each indicating a pairing of two study participants in a speed dating experiment event. Following cleaning, there were 543 observations, each representing a unique participant who took part in at least one speed dating event throughout the experiment.

Response Variable

Our response variable is the yes frequency of a participant amidst the New York City dating pool, which was created by a simple mutate operation involving dividing the number of matches received by a study participant by the total number of dates the participant took part in. This resulting continuous response variable was analyzed in conjunction with weighted explanatory variables discussed below, ultimately allowing for conclusion of what affects the appeal and general compatibility of a single on the market. However, we do acknowledge the limitations of the sample size of dates for each person may not be wholly representative of all singles, and the contexts within a metropolitan university population enacts a bias between populations of developed adults in varied geographies and states of life.

Explanatory Variables

In our analysis, we have chosen to include 39 explanatory variables, categorized into four groupings: Activities and Hobbies, Demographics, Preferences, and External Ratings. Detailed specification of these variables can be found in the appendix.

DATA: EXPLORATION: RESPONSE VARIABLE

To gain a better understanding of our response variable, we plotted its distribution, seen in **Figure 1** below. We used the full data instead of just the training set.

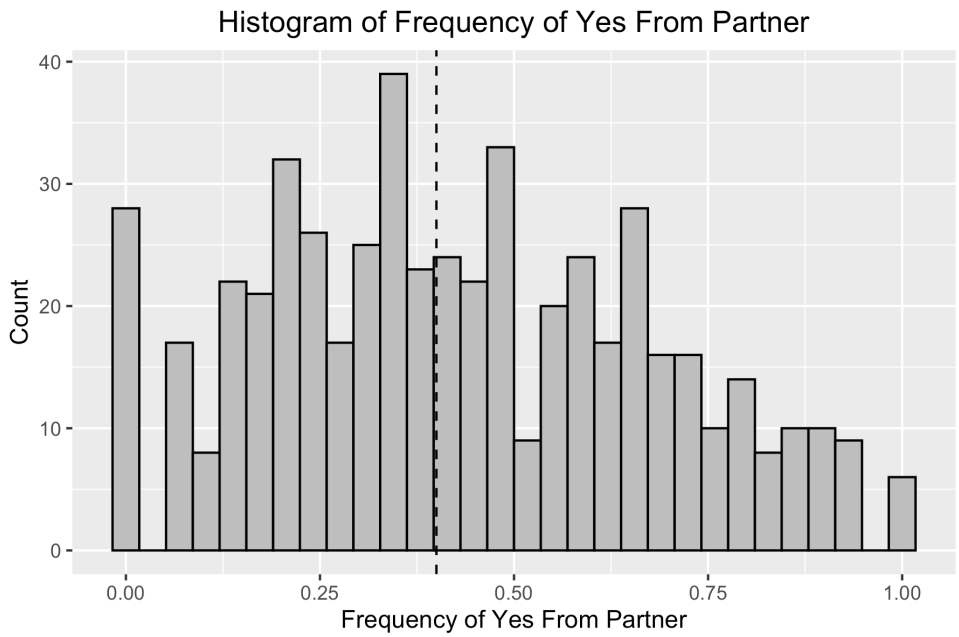


Figure 1: Histogram of Frequencies of Yes from Partners

This data is very slightly right-skewed as it has a slightly longer right tail, but it seems to have a slightly random, uniform distribution. The median frequency of yes from a partner is 0.4.

Additionally, we explored the relationship between some of our features and our response variable. In **Figure 2** below, we compared the distribution of the frequency of yes based on the subject's gender. As we can observe, females had a higher median frequency (0.5) than males (0.333), which suggests that women were far more likely to receive a yes from their partner than a man was.

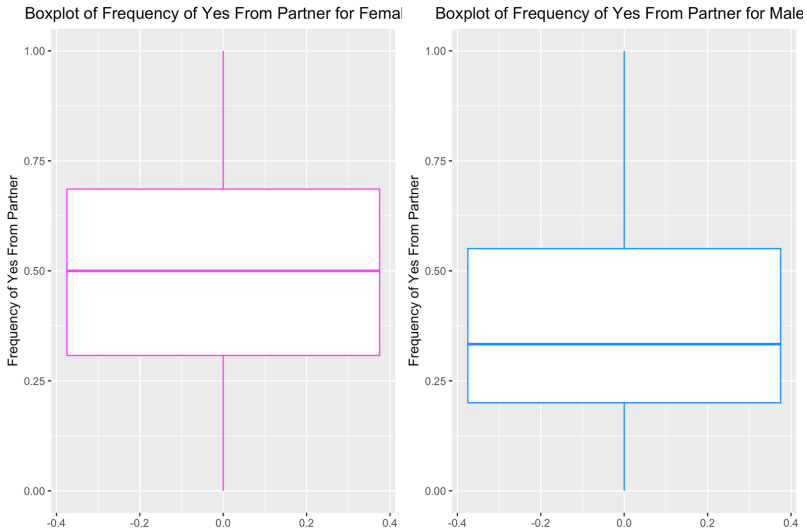


Figure 2: Side by Side Boxplots of Yes Frequencies from Partners, by Gender

We also explored the relationship between our response variable and race. **Table 1** below contains the mean frequency of yes from partners based on the subject's race. There were no Native Americans in the dataset despite it being a category for race in the dataset. We can see that Europeans/Caucasian-Americans had the highest frequency of yeses, while Asians/Pacific Islanders/Asian Americans had the lowest frequency. From this table, race does seem to have an impact on the frequency of yes from a subject's partners.

Table 1: Frequency of Yes, by Race

Race	Frequency of Yes
Black/African American	0.421
European/Caucasian-American	0.465
Latino/Hispanic American	0.476
Asian/Pacific Islander/Asian-American	0.356
Native American	NaN
Other	0.406

To garner preliminary insights into a relationship between our response variable and selected features, we first looked further into correlation between ratings of attractiveness, sincerity, intelligence, fun, ambition, and shared interest from all of a single participant's partners vs. yes frequency. As seen in **Figure 3** below, while all six attributes have a direct and positive relationship with the response variable, attractiveness, shared interest, and fun attributes show a much stronger relationship. This makes sense intuitively given the context of speed dating, as these are the traits that shine through most evidently within a four minute conversation with someone new, and shared interests particularly aid in comfort levels and desire to meet again to pursue a mutual love in hobby together.

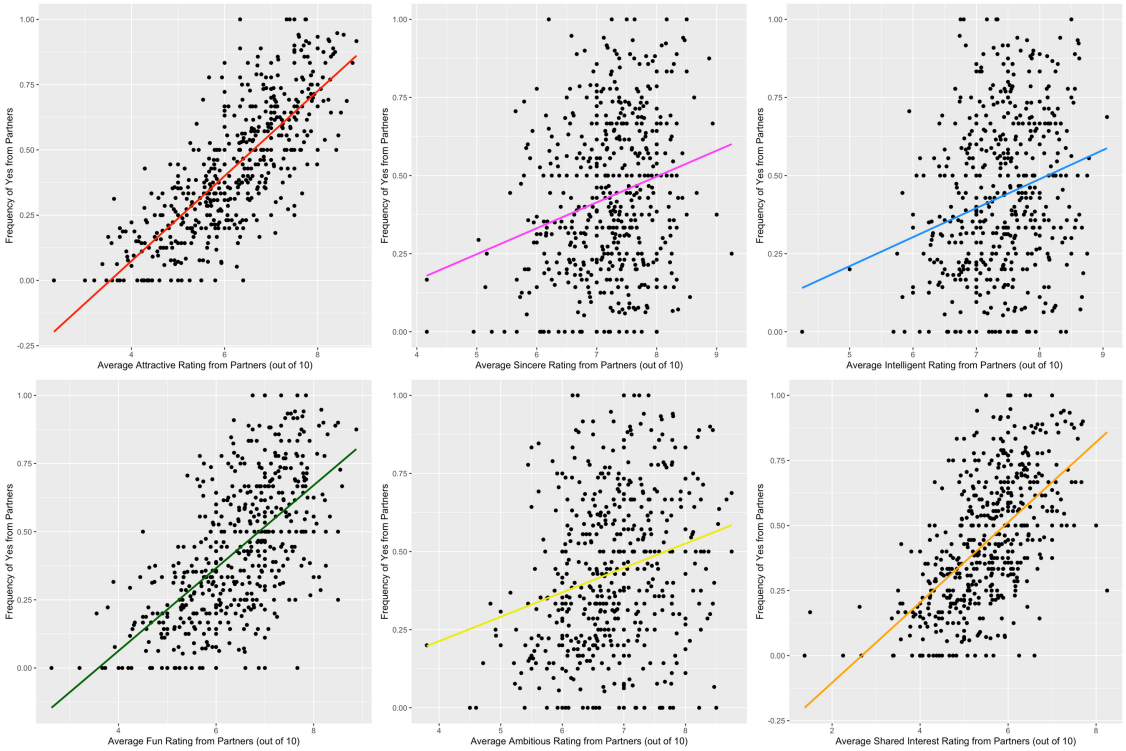
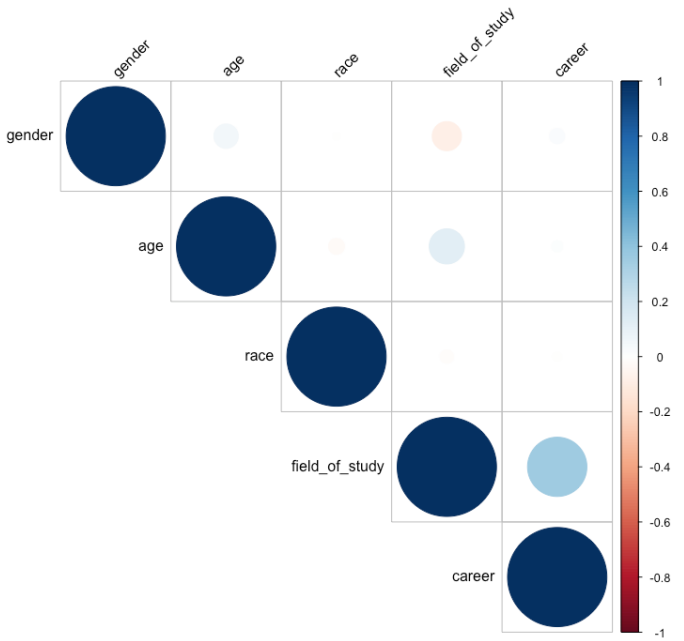


Figure 3: Frequency of Yes vs. Selected Attributes

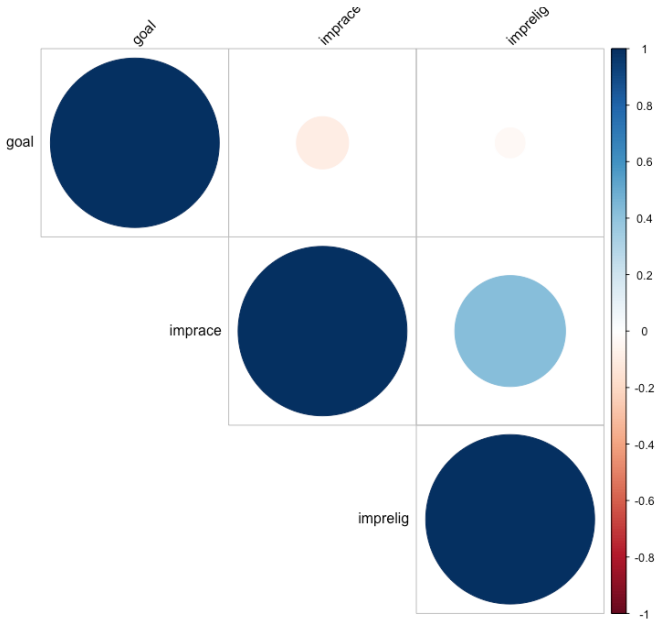
DATA: EXPLORATION: EXPLANATORY VARIABLES

Next, we explored the correlations of the explanatory, prediction variables with each other. We split our explanatory variables into four categories: demographics, dating preferences, activities and hobbies, and external feedback. The correlation plots are in **Figure 4** below:

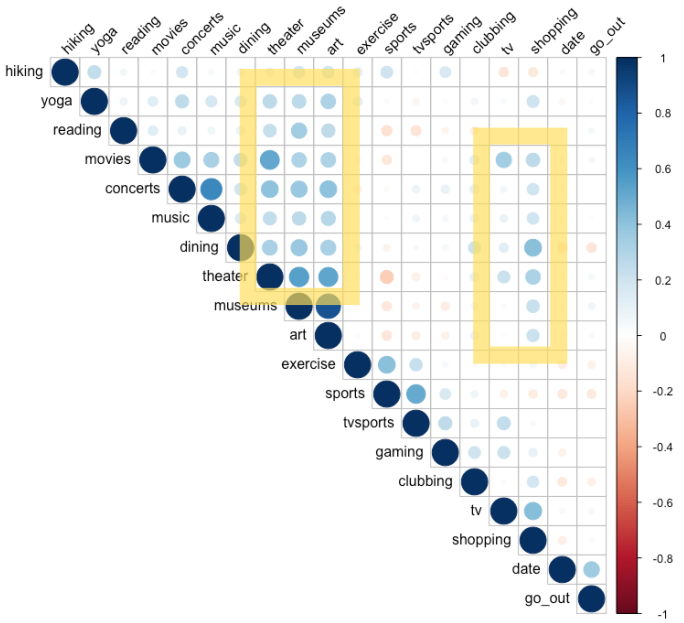
a. **Demographics:**



b. **Dating Preferences:**



c. **Activities and Hobbies:**



d. **External Feedback:**

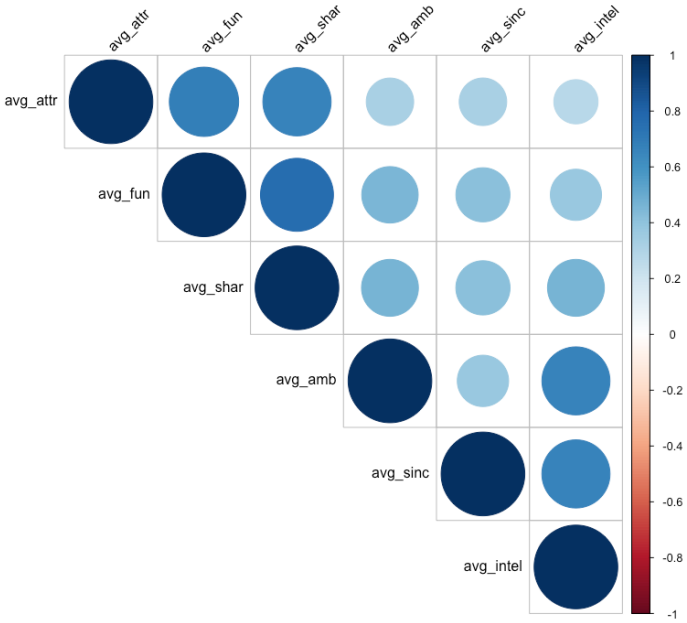


Figure 4: Correlation Plots of Explanatory Variables by Core Grouping:
a.) Demographics, b.) Preferences, c.) Activities and Hobbies, d.) External Feedback

Demographics: We do not observe many correlations between explanatory variables in the demographics category. Field of study and career are obviously positively correlated, which makes sense as one's field of study in college is inextricably linked to their career choice. Otherwise, there are not many strong positive or negative correlations, which means that our subjects did not have any underlying trends in their demographics, which is a good sign for our dataset.

Dating Preferences: We do not have many explanatory variables in the dating preferences category. However, we see a strong positive correlation between the importance of dating within one's religion and within one's race. This certainly makes sense. Race and religion are often linked factors (one's religious preferences depend on their race/culture), so it is reasonable to expect this correlation.

Activities and Hobbies: In this correlation plot, we see two clusters of positive correlations. The first, on the left, shows a positive correlation between a subject's preference for movies, concerts, music, dining, theater, museums, and arts. This is expected since these are all activities with similar audiences that typically involve some appreciation for art, culture, or food. On the right, we see that shopping also has positive correlations with movies, concerts, music, dining, theater, museums, and arts.

External Feedback: This correlation plot has a blue circle in every grid, implying that every explanatory variable in this category has a positive correlation with every other variable in this category. This is not necessarily to be expected, but it does make sense that if someone finds one characteristic of someone to be attractive, then they will also rate other characteristics of that person highly. Characteristics like attractiveness and intelligence are less positively correlated, which shows that physical and internal characteristics are less correlated. This correlation may cause confounding variables between the external feedback variables, but these correlations are still highly interesting to observe.

We also explored some of our explanatory variables on their own, as well as the relationship between pairs of explanatory variables.

The histogram below in **Figure 5** explores the distribution of the ages of the subjects in our dataset. The median age of our observations was exactly 26 years old. As expected, our age distribution is right-skewed, and our test subject population largely represents the young adult population.

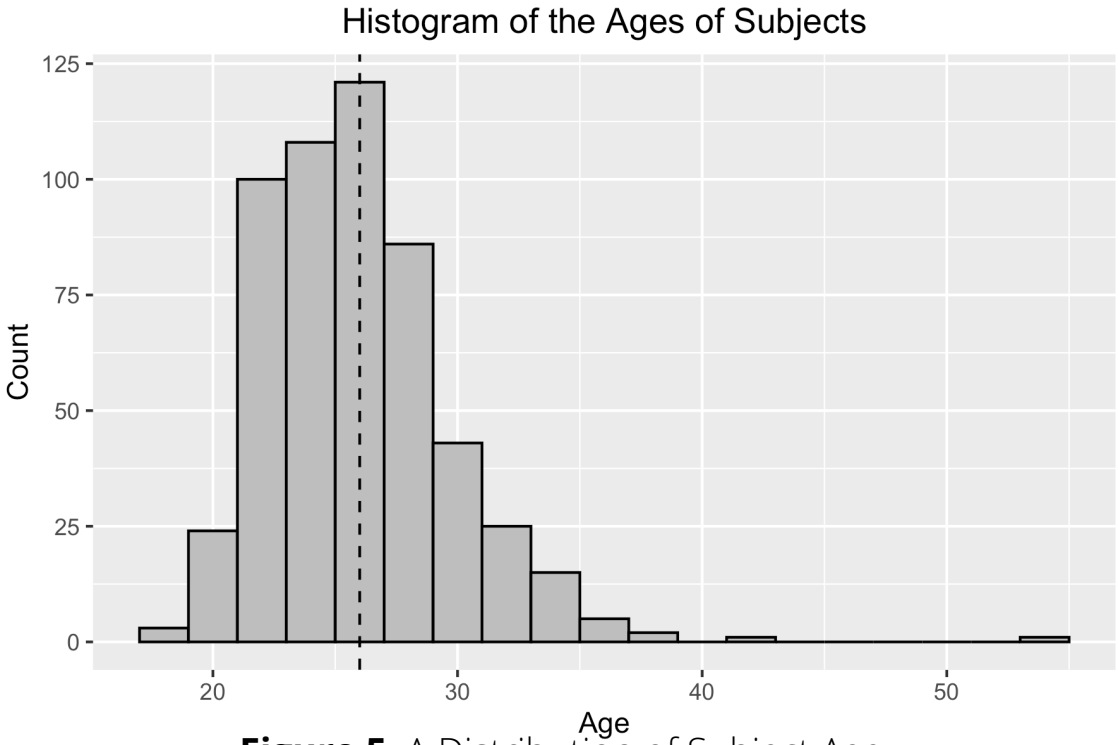


Figure 5: A Distribution of Subject Age

We also explored the relationship between gender and activity preferences. These results are in **Table 2** below. As seen in the Difference column, men, on average, strongly preferred playing sports, gaming, watching sports more than women. Women, on average, preferred shopping, theater, yoga, and art more than men.

Table 2: Difference in Preference for Hobby, by Gender

Activities	Female	Male	Difference
Sports	5.75	7.07	-1.321
TV Sports	4.10	4.96	-0.861
Exercise	6.41	6.20	0.204
Dining	8.18	7.41	0.773
Museum	7.51	6.53	0.978
Art	7.29	6.19	1.093
Hiking	5.94	5.56	0.371
Gaming	3.24	4.43	-1.189
Clubbing	5.92	5.58	0.337
Reading	7.93	7.40	0.529
TV	5.74	4.93	0.808
Theater	7.54	6.06	1.485
Movies	8.15	7.67	0.481
Concerts	7.16	6.56	0.606
Music	8.06	7.71	0.352
Shopping	6.53	4.73	1.794
Yoga	5.13	3.79	1.339

MODELING: REGRESSION METHODS

ORDINARY LEAST SQUARES

First, we ran an ordinary least squares regression of the yes frequency on all 39 of our explanatory variables. Full ordinary least squares regression results can be seen in the appendix. 5 of our explanatory variables have a significant relationship with our response variable at a 0.05 level: attraction rating (p-value significantly close to 0), sincerity rating (p-value of 0.0046), fun rating (p-value of 0.0092), gender (p-value of 0.0117), and shared interests rating (0.0303). This OLS regression has an r-squared value of 0.672, which means that our explanatory variables account for 67.2% of the variation in the response variable.

As seen in the appendix, the residual plot has no apparent correlation between fitted and residual plots, which is a good sign. Additionally, as seen in the appendix, the normal quantile plot is quite linear and does not deviate at the extreme values, which means that our data is apt for an OLS regression.

RIDGE AND LASSO REGRESSION

With an r-squared value of 0.647, the OLS regression seems to be quite fitting for the training data. However, to help minimize our variance, we built shrinkage regression models. Specifically, we built ridge and LASSO (Least Absolute Shrinkage and Selection Operator) cross-validated models. Cross-validation plots are in the appendix. Our ridge regression generated the following trace plot in **Figure 7,**

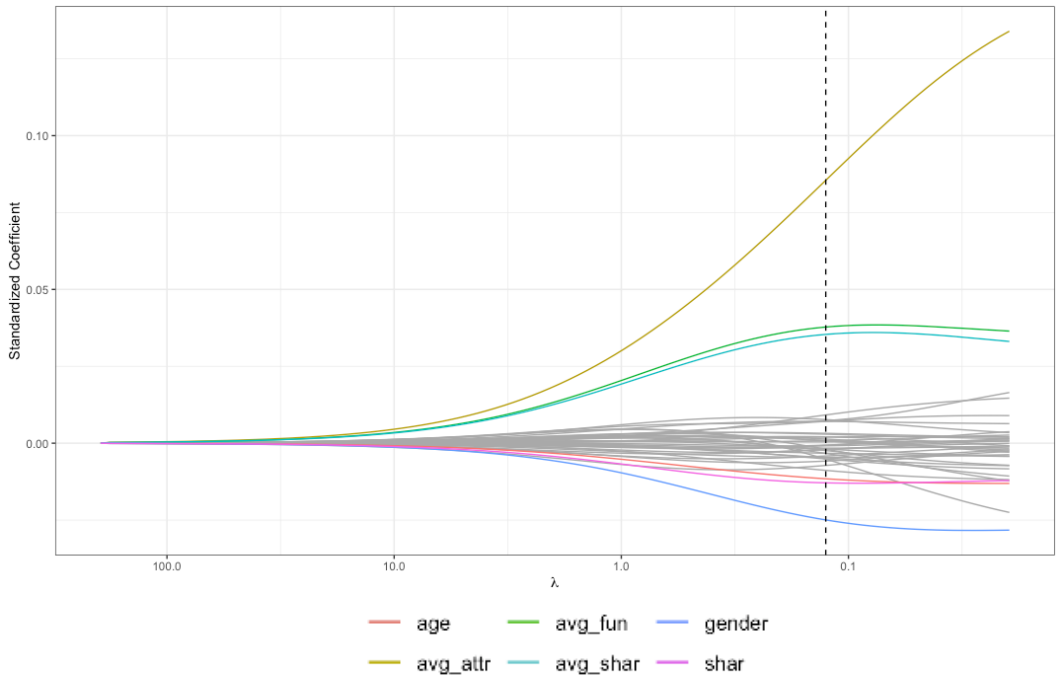


Figure 7: Trace Plot of Ridge Regression Model

Our trace plot highlighted the top 6 features, which were attractive rating, fun rating, shared interest rating, gender, shared interest preference, and age, in order of magnitude. Interestingly, as our penalty (lambda) decreases, the standardized coefficient for attractive rating seems to increase much faster than the other coefficients, demonstrating the weight in a person's attractiveness in getting a yes from a partner.

Meanwhile, LASSO regressions set some coefficients to zero, unlike the ridge regression. The LASSO regression produced the following trace plot in **Figure 8,**

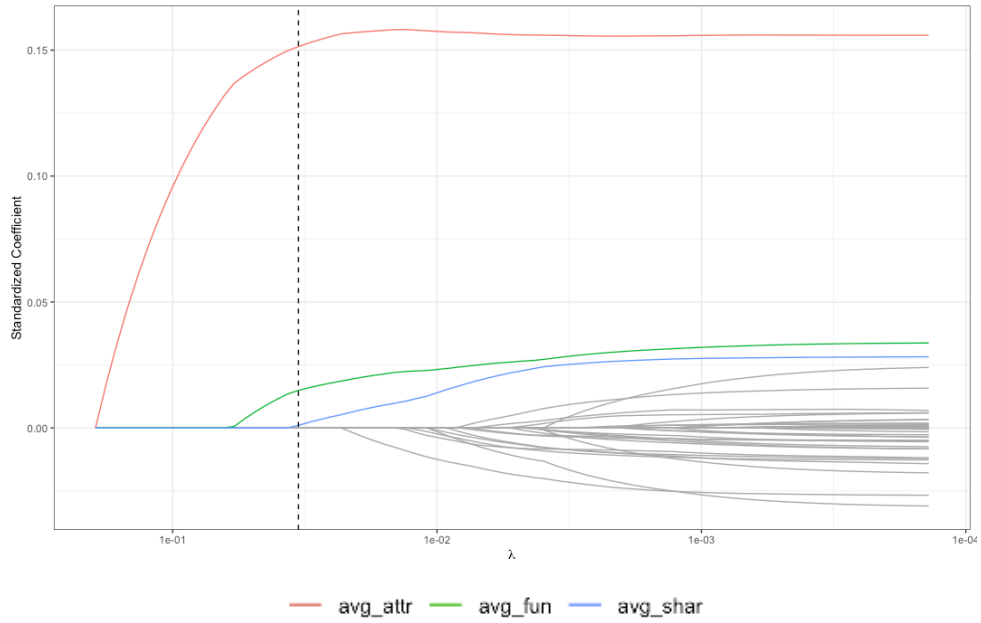


Figure 8: Trace Plot of LASSO Regression Model

At our optimal lambda, which was decided (by convention) by the one-standard-error rule, only three of our features have non-zero coefficients. These are, in order of magnitude, attractive rating, fun rating, and shared interest rating. These are the three most prevalent features found in the ridge regression, which is to be expected.
Both regression methods found that a partner's ratings for a person, as opposed to a person's interests or demographics, inform whether or not they would get a yes as a potential romantic partner. Specifically, a person's attractiveness, level of fun, and shared interests with another person are the most important criteria, and in an optimized LASSO regression, the only three features measured.

We chose these two regression methods over an elastic net, because running an elastic net gave us an alpha value of 1, which means it is equivalent to a LASSO fit. As we've found, there are a few features that very strongly inform the response, so it makes sense that we want to set some coefficients to zero to minimize the penalty, as is the case in a LASSO regression.

MODELING: TREE-BASED METHODS

PRUNED DECISION TREE

Next, to continue to model and understand our data, we ran tree-based methods. For this data, we produced regression trees, since we are trying to use these decision trees to predict at what frequency a person will receive a yes. To do this, we modeled an optimized decision tree, which helps us build a prediction model by setting conditional nodes that form a flowchart that each observation runs through.

First, we built a deepest possible tree, which contains too many fits and clearly overfits the data. However, we can prune the deepest tree by cutting off different branches and subtrees until we find something that optimizes the balance between complexity and simplicity; this is referred to as the bias-variance tradeoff and is key to building optimal models.

Once we tuned and optimized our decision tree, we produced the following in **Figure 9,**

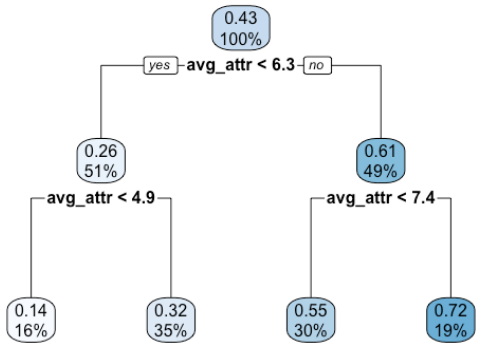


Figure 9: Pruned Decision Tree

This is quite the fascinating decision tree. When we optimize and prune it, the only relevant feature that node decisions rely on are attractive feature. Essentially, this tree suggests that the partner's rating for a person's attractiveness is the only significant indicator for how likely they are to get a yes.

MODELING: TREE-BASED METHODS

RANDOM FOREST

Next, we produced a random forest to improve our decision tree model. For this analysis, we used a default random forest, which runs a multitude of decision trees and averages and combines their results to provide a prediction. When we run our random forest, we are able to generate the follow variable importance plot in **Figure 10,**

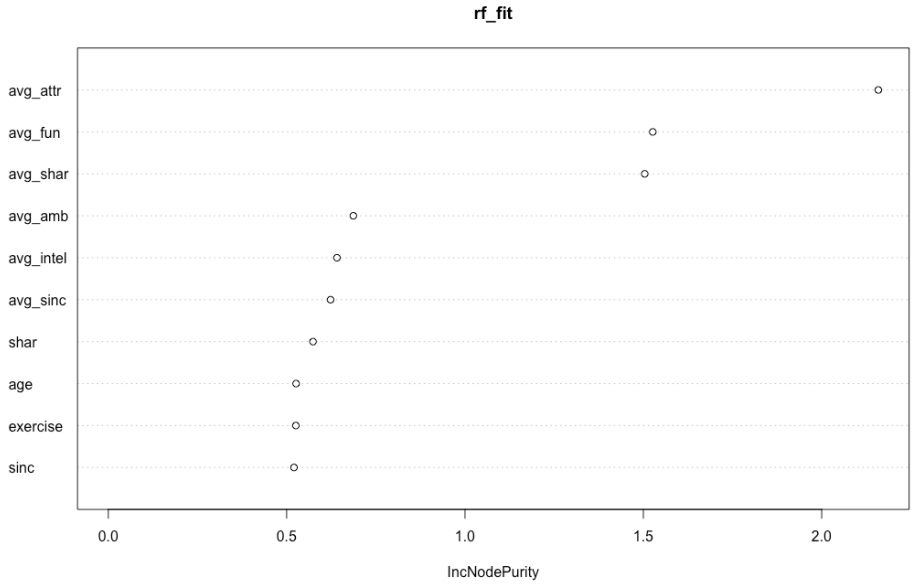


Figure 10: Variable Importance Plot, Random Forest Model

This Random Forest confirms much of what we have learned about the most important features in our model. Specifically, in our random forest, the 6 ratings given to a subject by their partner seemed to be the most important variables when determining the likelihood of getting a yes.

MODEL COMPARISON

We observe the following test residual mean squared errors (RMSE) in our models in **Table 3,**

Table 3: RMSE of All Models

Method	Train RMSE	Test RMSE
Ordinary Least-Squares	0.142	0.168
Ridge	0.152	0.171
Lasso	0.154	0.176
Decision Tree	0.315	0.190
Random Forest	0.106	0.208

CONCLUSION: PERFORMANCE AND LIMITATIONS

As the table above demonstrates, interestingly, our Ordinary Least-Squares regression model had the lowest test RMSE, while the Random Forest had the lowest training RMSE. The Ridge and LASSO regressions performed very similarly, which is reasonable, while the decision tree performed slightly worse on the test data. We can note some interesting observations here.

First, we see that the Random Forest was very accurate at modeling the training data, but not so much with the test data. This likely suggests severe overfitting with the random forest, which likely suggests that it is too complex for our data.

This hypothesis is also supported by the fact that our Ordinary Least-Squares Regression performed the best with our training data. This is by far the simplest regression method out of anything we used, which likely makes it the case that we do not need sophisticated machine learning methods to model this data.

This is evidenced by the fact that only a few features proved to be far better indicators than the rest. Typically, methods like Ridge, LASSO, decision trees, and random forests are meant for data with relatively more features compared with observations, which was not the case with our data, or data with many relevant features. We have since learned that this is likely not the case with this speed dating data, which likely supports our OLS regression model.

As we observed in our data exploration section, the attractive rating has a very strong linear relationship with yes frequency. Since this one feature was so dominant in all of our modeling methods, it then makes sense that our model as a whole is most accurate as a simple OLS regression.

Figure 11 below revisits this relationship between the six ratings each person was given and their yes frequency, and affirms the accuracy of our OLS regression model.

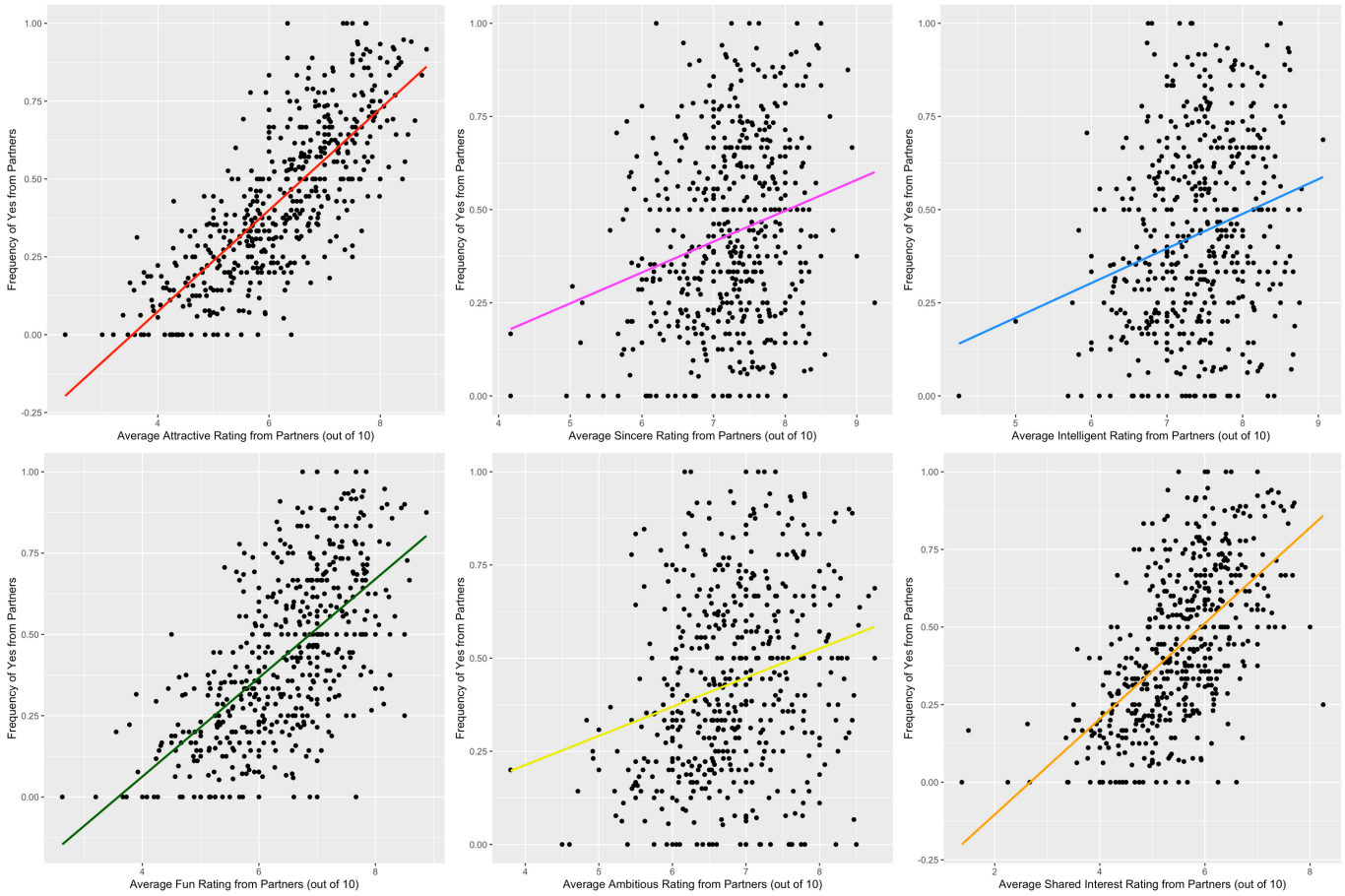


Figure 11: Ratings Granted to Participant vs. Response Variable
(L-R, T-D): Attractiveness, Sincerity, Intelligence, Fun, Ambitious, Shared Interest

CONCLUSION: LIMITATIONS

Within the data, it became clear that the context of speed dating, particularly the four minute time frame to meet, may have played a limiting role in allowing for accurate and thorough judgement of a participant's partner, specifically of attributes such as shared interests, ambition, and intelligence, accounting for limited correlation in the data exploration phase of the analysis. Additionally, it became apparent that not all races/ethnicities were equally represented within the data, which was intended to be representative of the study body the experiment was marketed to. This introduced difficulties in determining the correlation between race and other explanatory variables, as the sample size was often not large enough, particularly for Native Americans, who are not represented at all in the dataset. Following our background research and subsequent analysis, it is important to note that in addition to the 39 explanatory variables we have mentioned, there is a possibility that the incorporation of further factors including but not limited to future career path, current income, and more cosmetic and appearance information like eye color, height, weight, that are theoretically all considered but not specified under the realm of attractiveness.

CONCLUSION: RECOMMENDED FOLLOW-UP ANALYSES

To compensate for the aforementioned limitations, more up to date data may be used for further analyses on the same topic, expanding the target demographic well beyond the university sphere and incorporating a broader range of both ages and identities (racial, ethnic, sexual, etc.) to more accurately represent the current dating climate. Further, preliminary research may be conducted in a more free form manner, asking what singles in a selected area prioritize, then categorizing explanatory variables based on user response, garnering a more accurate selection of factors. Further, basic demographic and aesthetic factors may be asked upfront, as mentioned above in data limitations, to gain a better understanding of how participants value and interpret attractiveness, one of the most prominent factors found in the above analysis.

APPENDIX: EXPLANATORY VARIABLES

Denoted below are the 39 explanatory variables utilized in the analysis, with the variables names listed in parentheses.

Demographics

- **Gender (gender)**
  - Type: Categorical
  - Description: Gender of the participant, where female = 0, male = 1.

- **Age (age)**
  - Type: Continuous
  - Description: Age of the participant.

- **Field of study (field_of_study)**
  - Type: Categorical
  - Description: Field of the study of the participant, denoted by integers ranging from 1-18, where

    - 1 = Law
    - 2 = Math
    - 3= Social Science, psychologist
    - 4 = Medical science, pharmaceuticals, and bio tech
    - 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

APPENDIX: EXPLANATORY VARIABLES

- **Race (race)**
  - Type: Categorical
  - Description: Race of the participant, denoted by integers ranging from 1-6, where
    - 1 = Black/African American
    - 2 = European/Caucasian-American
    - 3 = Latino/Hispanic American
    - 4 = Asian/Pacific Islander/Asian-American
    - 5 = Native American
    - 6 = Other

- **Career (career)**
  - Type: Categorical
  - Description: Intended career of the participant, denoted by integers ranging from 1-17, where

    - 1 = Lawyer
    - 2 = Academic/Research
    - 3 = Psychologist
    - 4 = Doctor/Medicine
    - 5 =Engineer
    - 6 = Creative Arts/Entertainment
    - 7 = Banking/Consulting/Finance/Marketing/Business/CEO/Entrepreneur/ Admin
    - 8 = Real Estate
    - 9 = International/Humanitarian Affairs
    - 10 = Undecided
    - 11 = Social Work
    - 12 = Speech Pathology
    - 13 = Politics
    - 14 = Pro sports/Athletics
    - 15 = Other
    - 16 = Journalism
    - 17 = Architecture

APPENDIX: EXPLANATORY VARIABLES

Preferences

- **Importance of race (imprace)**
  - Type: Categorical
  - Description: The participant's relative weighting of importance that they date someone of the same racial and ethnic background, on a scale of 1-10.

- **Importance of religion (imprelig)**
  - Type: Categorical
  - Description: The participant's relative weighting of importance that they date someone of the same religious background, on a scale of 1-10.

- **Goal of of the date (goal)**
  - Type: Categorical
  - Description: The participant's ultimate goal from attending event(s) like these, denoted by integers 1-6, where

    - 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

Activities and Hobbies

- **Date frequency (date)**
  - Type: Categorical
  - Description: How frequently does the participant usually go out on dates, denoted by integers 1-7, where
    - 1 = Several times a week
    - 2 = Twice a week
    - 3 = Once a week
    - 4 = Twice a month
    - 5 = Once a month
    - 6 = Several times a year
    - 7 = Almost never

APPENDIX: EXPLANATORY VARIABLES

- **Importance of race (imprace)**
  - Type: Categorical
- **Outing frequency (go_out)**
  - Type: Categorical
  - Description: How frequently does the participant usually go out on dates, denoted by integers 1-7, where
    - 1 = Several times a week
    - 2 = Twice a week
    - 3 = Once a week
    - 4 = Twice a month
    - 5 = Once a month
    - 6 = Several times a year
    - 7 = Almost never
- **Interest in the following activities**, on a scale of 1-10 by the participant
  - **Sports (sports)**
    - Type: Categorical
    - Description: Interest in playing sports
  - **Watching sports (tvsports)**
    - Type: Categorical
    - Description: Watching sports on TV
  - **Exercise (exercise)**
    - Type: Categorical
    - Description: Working out, exercising, and body building
  - **Dining (dining)**
    - Type: Categorical
    - Description: Going out to dine, enjoying meals outside of the house
  - **Museums (museums)**
    - Type: Categorical
    - Description: Attending/exploring museums and galleries
  - **Art (art)**
    - Type: Categorical
    - Description: Enjoyment of learning about and immersing oneself in art exploration
  - **Hiking (hiking)**
    - Type: Categorical
    - Description: Going hiking or camping

APPENDIX: EXPLANATORY VARIABLES

Interest in the following activities, on a scale of 1-10 by the participant, cont.

- **Gaming (gaming)**
  - Type: Categorical
  - Description: Gaming of any form
- **Clubbing (clubbing)**
  - Type: Categorical
  - Description: Dancing and/or clubbing, enjoyment of the clubbing scene
- **Reading (reading)**
  - Type: Categorical
  - Description: Enjoyment of reading
- **TV (tv)**
  - Type: Categorical
  - Description: Watching TV
- **Theater**
  - Type: Categorical
  - Description: Enjoyment of theater and theater arts
- **Movies (movies)**
  - Type: Categorical
  - Description: talking about, enjoying, learning more about movies
- **Concerts (concerts)**
  - Type: Categorical
  - Description: attending concerts for various artists
- **Music (music)**
  - Type: Categorical
  - Description: talking about and listening to music
- **Shopping (shopping)**
  - Type: Categorical
  - Description: enjoyment of shopping
- **Yoga (yoga)**
  - Type: Categorical
  - Description: enjoyment of yoga

APPENDIX: EXPLANATORY VARIABLES

Relative importance of each of the following: Participants were given 100 points and asked to distribute them between the following six categories, indicating more points for attributes which are deemed more important in a potential dates.

1. **Attractiveness (attr)**
2. **Sincerity (sinc)**
3. **Intelligence (intel)**
4. **Fun (fun)**
5. **Ambition (amb)**
6. **Shared Interests (shar)**

External Feedback

Participants were rated by all of their dates on a scale of 1-10, with 10 being the most of the selected trait and 1 denoted that the participant was lacking in this area. The values were then averaged across all dates for the selected participant, and the six attributes in this category were as follows:

1. **Attractiveness (avg_attr)**
2. **Sincerity (avg_sinc)**
3. **Intelligence (avg_intel)**
4. **Fun (avg_fun)**
5. **Ambition (avg_amb)**
6. **Shared Interests (avg_shar)**

APPENDIX: ORDINARY LEAST SQUARES

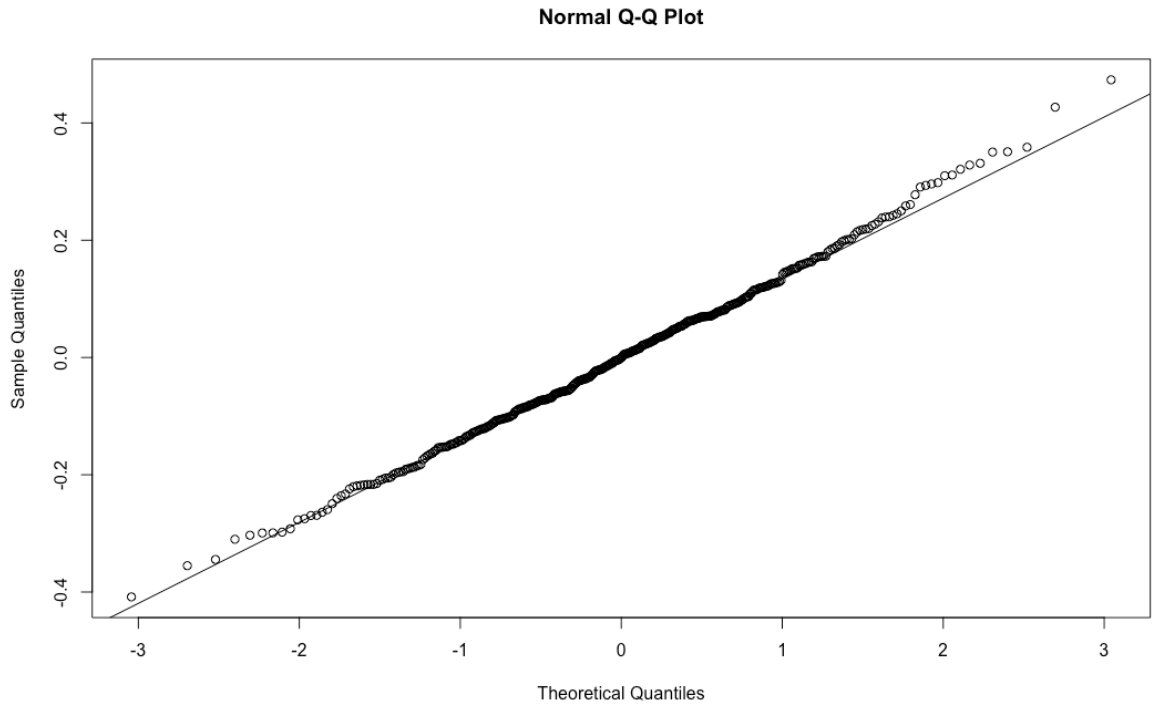


Figure 12: Normal QQ Plot

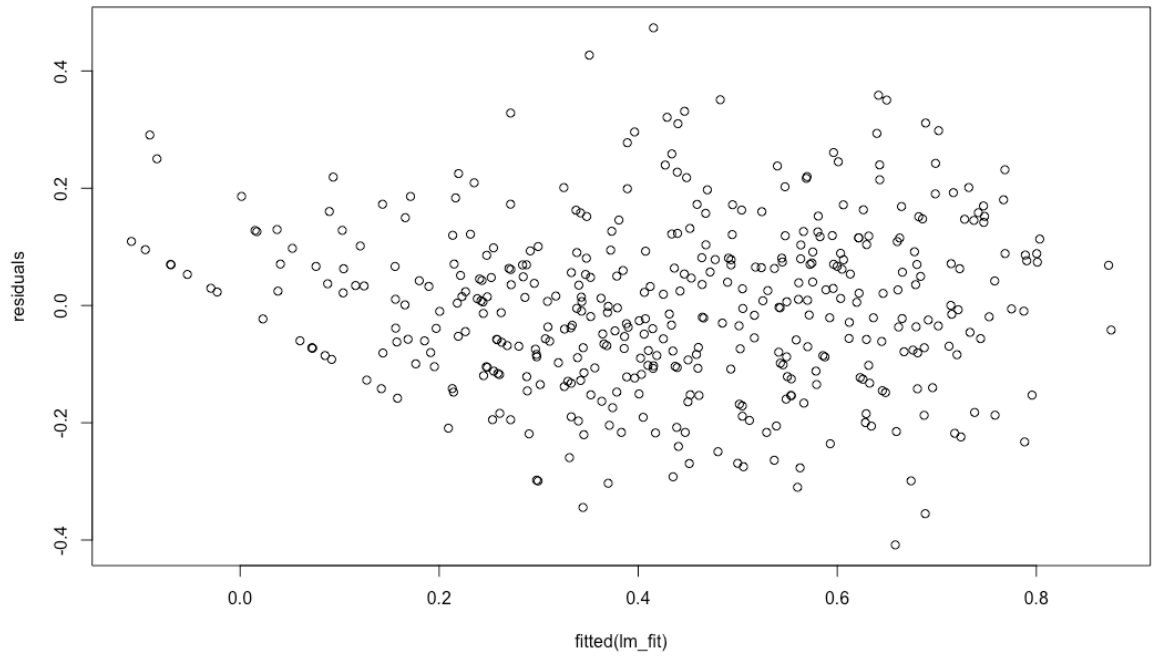


Figure 13: Residual Plot

APPENDIX: CROSS VALIDATION PLOTS

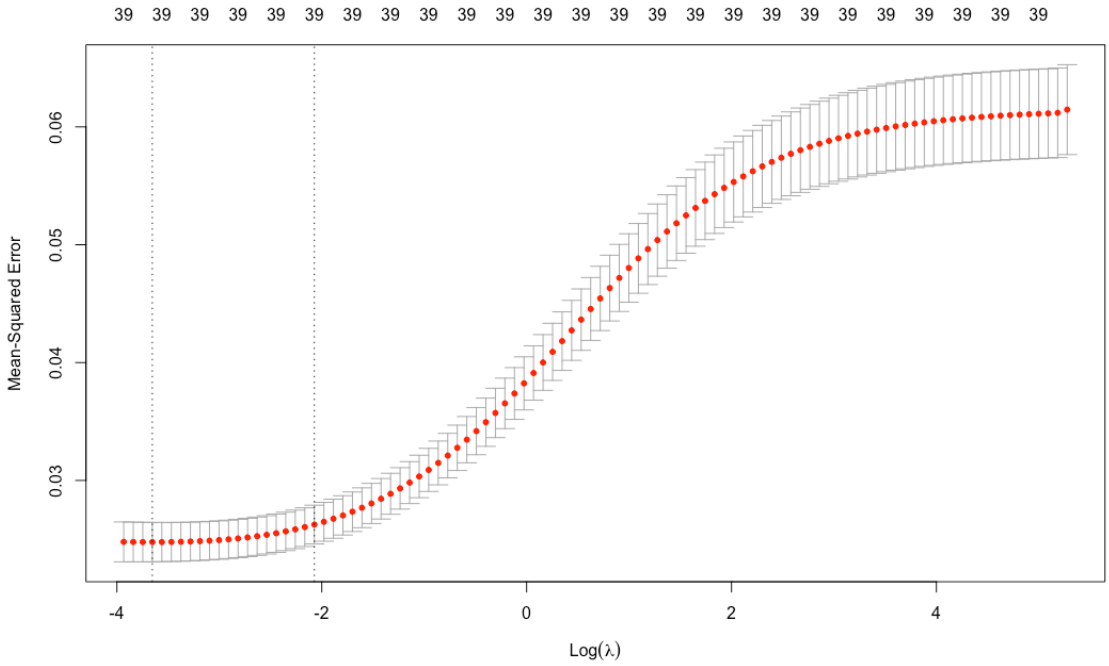


Figure 14: Ridge CV Plot

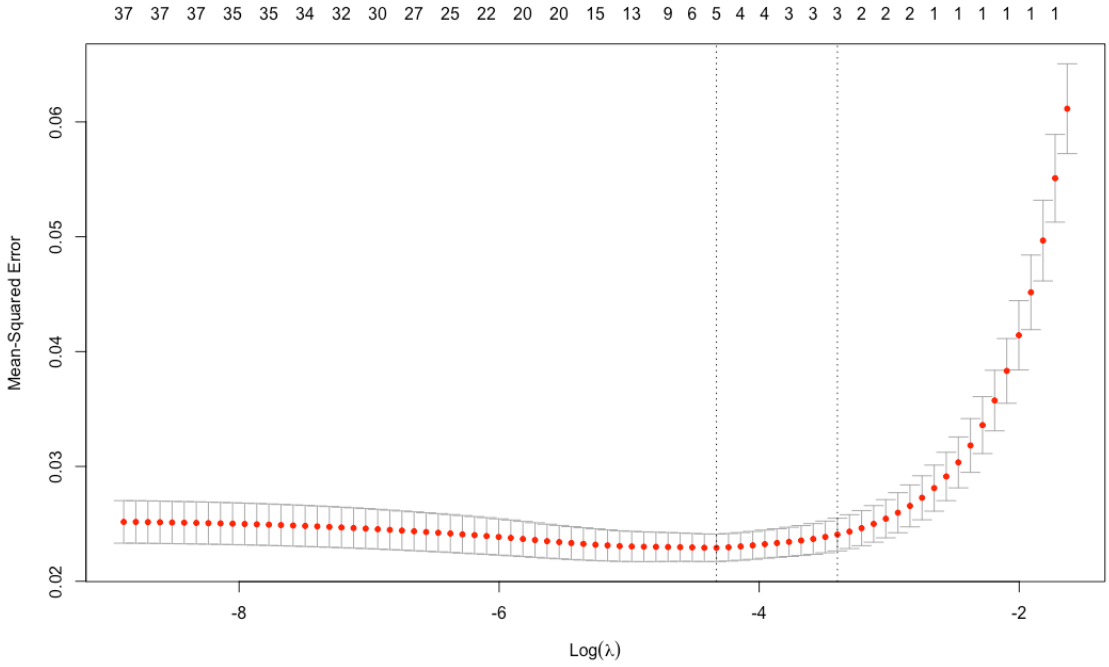


Figure 15: Lasso CV Plot