University of Waterloo

DON'T DRINK AND DATE

A Study in Speed Dating

Final Project Report

For

STAT 841 Statistical Learning: Classification (Fall 2019)

Submitted By

Team: gSpSmS

Govind Sharma (20817244)

Puneeth Srinivas Mohan Saladi (20835628)

Contents

Motivation

Introduction

The whole world is moving at a brisk pace with so much competition, responsibilities and distractions. This trend has been prevalent ever since the introduction of capitalism and it will only continue to grow in the future. As such, people have to make compromises on their personal lives, take away from their social circles and be more and more consumed in this never ending race.

In such an environment, the process of seeking a partner for a romantic relationship is even more daunting a task than it already was. People are often finding it hard to make time for this very important aspect of life. But as is the case with many other situations, with new problems come new solutions. And that is why we continue to see a surge in various means of facilitating the "dating process". Be it online portals or mobile applications, there always seems to be the "next new thing" in the dating industry.

We are going to focus on one such relatively old but still very popular form of organized romance: speed dating. Speed dating is a formalized matchmaking process which has the purpose of encouraging eligible singles to meet large numbers of new potential partners in a very short period of time. These types of arrangements are particularly popular in western societies where polygamous relationships are a common practice.

Speed dating is useful not only because it creates an environment conducive for people to find multiple partners in a very short period of time but it is also interesting from a scientific point of view. Speed dating can be viewed as an experiment in human interaction and data collected from this experiment can be analyzed in a variety of ways to conclude many useful things, patterns and behaviours that govern the fundamentals of partner selection in humans.

Problem Statements

Using a dataset (described in subsequent sections) and after applying various tools for analysis, visualization and prediction, we would like to have some insights regarding the following questions:

1. Given a set of attributes pertaining to potential partners, can we accurately predict whether two people will be ready to date each other?
2. Based on a person's own preferences, and the attributes of the potential partner, will the person be interested in the opposite person?
3. For each gender (male & female), which attributes are most desirable in their respective partners?
4. For people of different races, are there some qualities that outshine others when it comes to choosing a partner?
5. How does a person's preferences change with age?

Data

Description

We will be using data which was collected from a project named "**Speed Dating Experiment**".[3] This experiment was conducted by **Professor Ray Fisman** and **Professor Sheena Iyengar** from Columbia Business School. They introduced this data set originally as part of their paper, "**Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment.**"[1] As described in the paper, this data was gathered from people who volunteered for a speed dating event that spanned from the time period of 2002-2004.

Here is how the experiment was conducted:

In the event, the participants were given a chance to have 4 minutes of "first interaction" with all other potential partners. After every such interaction, each participant will rate the person they just met based on various attributes. Together with rating each person gives for various attributes, the participants will also write down a final verdict on whether or not they will be willing to go on a second "formal" date. The attributes on which each person is rated by their partner are, "**attractiveness**", "**sincerity**", "**intelligence**", "**fun**", "**ambition**" and "**shared interests**". Another thing to note here is that each person has also rated themselves on all of these attributes before meeting anyone else. So the database contains data from both perspectives.

The data set also includes data on various preferences a person has for their potential partners. For example, how important is it for a person that their partner be of the same race, etc. This along with all other attributes in the dataset are summarized below:

Attribute	**Description**
gender	Gender of self
age	Age of self
age_o	Age of partner
d_age	Difference in age
race	Race of self
race_o	Race of partner
samerace	Whether the two people have the same race
importance_same_race	How important is it for a person to have the partner be of the same race
importance_same_religion	How important is it for a person to have the partner be of the same religion
field	person's field of work/study
pref_o_attractiveness	How important does partner rate attractiveness
pref_o_sincere	How important does partner rate sincerity
pref_o_intelligence	How important does partner rate intelligence
pref_o_funny	How important does partner rate being funny
pref_o_ambitious	How important does partner rate ambition
pref_o_shared_interests	How important does partner rate having shared interests
attractive_o	Rating by partner at night of event on attractiveness
sincere_o	Rating by partner at night of event on sincerity
intelligence_o	Rating by partner at night of event on intelligence
funny_o	Rating by partner at night of event on being funny
ambitious_o	Rating by partner at night of event on being ambitious
shared_interests_o	Rating by partner (about me) at night of event on shared interest
attractive_important	What do you look for in a partner - attractiveness
sincere_important	What do you look for in a partner - sincerity
intelligence_important	What do you look for in a partner - intelligence
funny_important	What do you look for in a partner - being funny
ambitious_important	What do you look for in a partner - ambition
shared_important_important	What do you look for in a partner - shared interests

attractive	Rate yourself - attractiveness
sincere	Rate yourself - sincerity
intelligence	Rate yourself - intelligence
funny	Rate yourself - being funny
ambition	Rate yourself - ambition
attractive_partner	Rate your partner - attractiveness
sincerity_partner	Rate your partner - sincerity
intelligence_partner	Rate your partner - intelligence
funny_partner	Rate your partner - being funny
ambition_partner	Rate your partner - ambition
sports	Your own interests [1-10]
tvsports	Your own interests [1-10]
exercise	Your own interests [1-10]
dining	Your own interests [1-10]
museums	Your own interests [1-10]
art	Your own interests [1-10]
hiking	Your own interests [1-10]
gaming	Your own interests [1-10]
clubbing	Your own interests [1-10]
reading	Your own interests [1-10]
tv	Your own interests [1-10]
theater	Your own interests [1-10]
movies	Your own interests [1-10]
concerts	Your own interests [1-10]
music	Your own interests [1-10]
shopping	Your own interests [1-10]
yoga	Your own interests [1-10]
interests_correlate	Correlation between participant's and partner's ratings of interests
expected_happy_with_sd_people	How happy do you expect to be with the people at during event?
expected_num_interested_in_me	Out of the 20, how many do you expect will be interested in you?
expected_num_matches	How many matches do you expect to get?
like	Did you like your partner?
guess_prob_liked	How likely do you think it is that your partner likes you?
met	Have you met your partner before?
decision	Decision at night of event.
decision_o	Decision of partner at night of event.
match	Match (yes/no)

Preprocessing

As the dataset had a lot duplicate fields and missing data, some amount of preprocessing was required to be performed. We performed data cleaning, data transformation, and data reduction as part of preprocessing. Firstly, the dataset had a number of features which were basically bucketed categorical variables of other present features. These features have been listed below:

d_d_age	d_importance_same_race	d_importance_same_religion
d_pref_o_attractive	d_pref_o_sincere	d_pref_o_intelligence
d_pref_o_funny	d_pref_o_ambitious	d_pref_o_shared_interests
d_attractive_o	d_sincere_o	d_intelligence_o

d_funny_o	d_ambitious_o	d_shared_interests_o
d_attractive_important	d_sincere_important	d_intelligence_important
d_funny_important	d_ambitious_important	d_shared_interests_important
d_attractive	d_sincere	d_intelligence
d_funny	d_ambitious	d_shared_interests
d_attractive_partner	d_sincere_partner	d_intelligence_partner
d_funny_partner	d_ambitious_partner	d_shared_interests_partner
d_sports	d_tvsports	d_exercise
d_dining	d_museums	d_art
d_hiking	d_gaming	d_clubbing
d_reading	d_tv	d_theater
d_movies	d_concerts	d_music
d_shopping	d_yoga	d_interests_correlate
d_expected_happy_with_sd_people	d_expected_num_interested_in_me	d_expected_num_matches
d_like	d_guess_prob_liked	

All of these fields were removed. Along with this, two more fields - has_null and wave, which represented if the observation has a null value and the round of speed dating respectively were also removed. After this, all the fields were converted to to either numeric or factor as required. Although many of the machine learning models require the predictors to be all numeric, this task was left to be done during the time of model creation.

Next we had to deal with missing values for both numeric and categorical type. Rather then using some existing library, which handles the imputation, we decided to fill in the missing values manually with mean and mode for numeric and categorical variables respectively. After the missing values are filled, we crosscheck if any more missing values exist.

Descriptive Analysis

Before we commence the predictive analysis for the given dataset, we have to do some analysis of the dataset as it exists. There is a lot of information and conclusions that can be drawn just by looking closely at the data.

Let's look at the summary of the data at first:

```
##     gender         age            age_o           d_age       
##  female:4184   Min.   :18.00   Min.   :18.00   Min.   : 0.000  
##  male  :4194   1st Qu.:24.00   1st Qu.:24.00   1st Qu.: 1.000  
##                Median :26.00   Median :26.00   Median : 3.000  
##                Mean   :26.36   Mean   :26.36   Mean   : 4.186  
##                3rd Qu.:28.00   3rd Qu.:28.00   3rd Qu.: 5.000  
##                Max.   :55.00   Max.   :55.00   Max.   :37.000  
##                                                                
##                                        race     
##  'Asian/Pacific Islander/Asian-American':1982  
##  'Black/African American'               : 420  
##  'Latino/Hispanic American'             : 664  
##  European/Caucasian-American            :4790  
##  Other                                  : 522  
##                                                
##                                                
##                                       race_o     samerace
##  'Asian/Pacific Islander/Asian-American':1978   0:5062  
##  'Black/African American'               : 420   1:3316  
##  'Latino/Hispanic American'             : 664           
##  European/Caucasian-American            :4795           
##  Other                                  : 521           
##                                                         
##                                                         
##  importance_same_race importance_same_religion
##  Min.   : 0.000       Min.   : 1.000          
##  1st Qu.: 1.000       1st Qu.: 1.000          
##  Median : 3.000       Median : 3.000          
##  Mean   : 3.785       Mean   : 3.652          
##  3rd Qu.: 6.000       3rd Qu.: 6.000          
##  Max.   :10.000       Max.   :10.000          
##                                               
##                        field      pref_o_attractive pref_o_sincere
##  business                : 694   Min.   :  0.0     Min.   : 0.0  
##  law                     : 604   1st Qu.: 15.0     1st Qu.:15.0  
##  mba                     : 468   Median : 20.0     Median :18.0  
##  'social work'           : 414   Mean   : 22.5     Mean   :17.4  
##  'international affairs' : 287   3rd Qu.: 25.0     3rd Qu.:20.0  
##  'electrical engineering': 223   Max.   :100.0     Max.   :60.0  
##  (Other)                 :5688                                   
##  pref_o_intelligence  pref_o_funny   pref_o_ambitious
##  Min.   : 0.00       Min.   : 0.00   Min.   : 0.00   
##  1st Qu.:17.65       1st Qu.:15.00   1st Qu.: 5.00   
##  Median :20.00       Median :18.00   Median :10.00   
##  Mean   :20.27       Mean   :17.46   Mean   :10.69   
##  3rd Qu.:23.26       3rd Qu.:20.00   3rd Qu.:15.00   
##  Max.   :50.00       Max.   :50.00   Max.   :53.00   
```

```
##
##  pref_o_shared_interests  attractive_o     sincere_o      intelligence_o
##  Min.   : 0.00           Min.   : 0.00   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:10.00           1st Qu.: 5.00   1st Qu.: 6.000   1st Qu.: 7.000
##  Median :11.36           Median : 6.00   Median : 7.000   Median : 7.369
##  Mean   :11.85           Mean   : 6.19   Mean   : 7.175   Mean   : 7.369
##  3rd Qu.:15.69           3rd Qu.: 8.00   3rd Qu.: 8.000   3rd Qu.: 8.000
##  Max.   :30.00           Max.   :10.50   Max.   :10.000   Max.   :10.000
##
##    funny_o        ambitious_o     shared_interests_o attractive_important
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000     Min.   :  0.00
##  1st Qu.: 5.000   1st Qu.: 6.000   1st Qu.: 4.000     1st Qu.: 15.00
##  Median : 6.401   Median : 7.000   Median : 5.475     Median : 20.00
##  Mean   : 6.401   Mean   : 6.778   Mean   : 5.475     Mean   : 22.51
##  3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 7.000     3rd Qu.: 25.00
##  Max.   :11.000   Max.   :10.000   Max.   :10.000     Max.   :100.00
##
##  sincere_important intelligence_important funny_important
##  Min.   : 0.00     Min.   : 0.00          Min.   : 0.00
##  1st Qu.:15.00     1st Qu.:17.65          1st Qu.:15.00
##  Median :18.18     Median :20.00          Median :18.00
##  Mean   :17.40     Mean   :20.27          Mean   :17.46
##  3rd Qu.:20.00     3rd Qu.:23.26          3rd Qu.:20.00
##  Max.   :60.00     Max.   :50.00          Max.   :50.00
##
##  ambitious_important shared_interests_important   attractive
##  Min.   : 0.00       Min.   : 0.00              Min.   : 2.000
##  1st Qu.: 5.00       1st Qu.:10.00              1st Qu.: 6.000
##  Median :10.00       Median :11.11              Median : 7.000
##  Mean   :10.68       Mean   :11.85              Mean   : 7.085
##  3rd Qu.:15.00       3rd Qu.:15.69              3rd Qu.: 8.000
##  Max.   :53.00       Max.   :30.00              Max.   :10.000
##
##    sincere        intelligence        funny          ambitious
##  Min.   : 2.000   Min.   : 2.000   Min.   : 3.000   Min.   : 2.000
##  1st Qu.: 8.000   1st Qu.: 7.000   1st Qu.: 8.000   1st Qu.: 7.000
##  Median : 8.147   Median : 8.000   Median : 8.000   Median : 8.000
##  Mean   : 8.295   Mean   : 7.704   Mean   : 8.404   Mean   : 7.578
##  3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.: 9.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##  attractive_partner sincere_partner  intelligence_partner funny_partner
##  Min.   : 0.00      Min.   : 0.000   Min.   : 0.000       Min.   : 0.000
##  1st Qu.: 5.00      1st Qu.: 6.000   1st Qu.: 7.000       1st Qu.: 5.000
##  Median : 6.00      Median : 7.000   Median : 7.369       Median : 6.401
##  Mean   : 6.19      Mean   : 7.175   Mean   : 7.369       Mean   : 6.401
##  3rd Qu.: 8.00      3rd Qu.: 8.000   3rd Qu.: 8.000       3rd Qu.: 8.000
##  Max.   :10.00      Max.   :10.000   Max.   :10.000       Max.   :10.000
##
##  ambitious_partner shared_interests_partner     sports
##  Min.   : 0.000    Min.   : 0.000           Min.   : 1.000
##  1st Qu.: 6.000    1st Qu.: 4.000           1st Qu.: 5.000
##  Median : 7.000    Median : 5.475           Median : 7.000
##  Mean   : 6.778    Mean   : 5.475           Mean   : 6.425
##  3rd Qu.: 8.000    3rd Qu.: 7.000           3rd Qu.: 9.000
##  Max.   :10.000    Max.   :10.000           Max.   :10.000
##
##     tvsports         exercise          dining          museums
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 5.000   1st Qu.: 7.000   1st Qu.: 6.000
##  Median : 4.000   Median : 6.246   Median : 8.000   Median : 7.000
##  Mean   : 4.575   Mean   : 6.246   Mean   : 7.784   Mean   : 6.986
##  3rd Qu.: 7.000   3rd Qu.: 8.000   3rd Qu.: 9.000   3rd Qu.: 8.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##       art              hiking           gaming          clubbing
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 5.000   1st Qu.: 4.000   1st Qu.: 2.000   1st Qu.: 4.000
##  Median : 7.000   Median : 6.000   Median : 3.000   Median : 6.000
##  Mean   : 6.715   Mean   : 5.737   Mean   : 3.881   Mean   : 5.746
##  3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 6.000   3rd Qu.: 8.000
##  Max.   :10.000   Max.   :10.000   Max.   :14.000   Max.   :10.000
##
##     reading             tv            theater           movies
##  Min.   : 1.000   Min.   : 1.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 7.000   1st Qu.: 3.000   1st Qu.: 5.000   1st Qu.: 7.00
##  Median : 8.000   Median : 6.000   Median : 7.000   Median : 8.00
##  Mean   : 7.679   Mean   : 5.304   Mean   : 6.776   Mean   : 7.92
##  3rd Qu.: 9.000   3rd Qu.: 7.000   3rd Qu.: 8.000   3rd Qu.: 9.00
##  Max.   :13.000   Max.   :10.000   Max.   :10.000   Max.   :10.00
##
##     concerts          music          shopping           yoga
##  Min.   : 0.000   Min.   : 1.000   Min.   : 1.000   Min.   : 0.000
##  1st Qu.: 5.000   1st Qu.: 7.000   1st Qu.: 4.000   1st Qu.: 2.000
##  Median : 7.000   Median : 8.000   Median : 6.000   Median : 4.000
##  Mean   : 6.825   Mean   : 7.851   Mean   : 5.631   Mean   : 4.339
##  3rd Qu.: 8.000   3rd Qu.: 9.000   3rd Qu.: 8.000   3rd Qu.: 6.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##  interests_correlate expected_happy_with_sd_people
##  Min.   :-0.830      Min.   : 1.000
##  1st Qu.:-0.010      1st Qu.: 5.000
##  Median : 0.200      Median : 6.000
##  Mean   : 0.196      Mean   : 5.534
##  3rd Qu.: 0.430      3rd Qu.: 7.000
##  Max.   : 0.910      Max.   :10.000
##
##  expected_num_interested_in_me expected_num_matches      like
##  Min.   : 0.000                Min.   : 0.000       Min.   : 0.000
##  1st Qu.: 5.571                1st Qu.: 2.000       1st Qu.: 5.000
##  Median : 5.571                Median : 3.000       Median : 6.000
##  Mean   : 5.571                Mean   : 3.208       Mean   : 6.134
##  3rd Qu.: 5.571                3rd Qu.: 4.000       3rd Qu.: 7.000
##  Max.   :20.000                Max.   :18.000       Max.   :10.000
##
##  guess_prob_liked      met          decision decision_o match
##  Min.   : 0.000   Min.   :0.00000   0:4860   0:4863     0:6998
##  1st Qu.: 4.000   1st Qu.:0.00000   1:3518   1:3515     1:1380
##  Median : 5.000   Median :0.00000
##  Mean   : 5.208   Mean   :0.04986
##  3rd Qu.: 7.000   3rd Qu.:0.00000
##  Max.   :10.000   Max.   :8.00000
##
```

As we can see all the attributes are properly scaled after preprocessing. All missing values have been dealt with and our data is ready to be used.

Ethnic Distribution

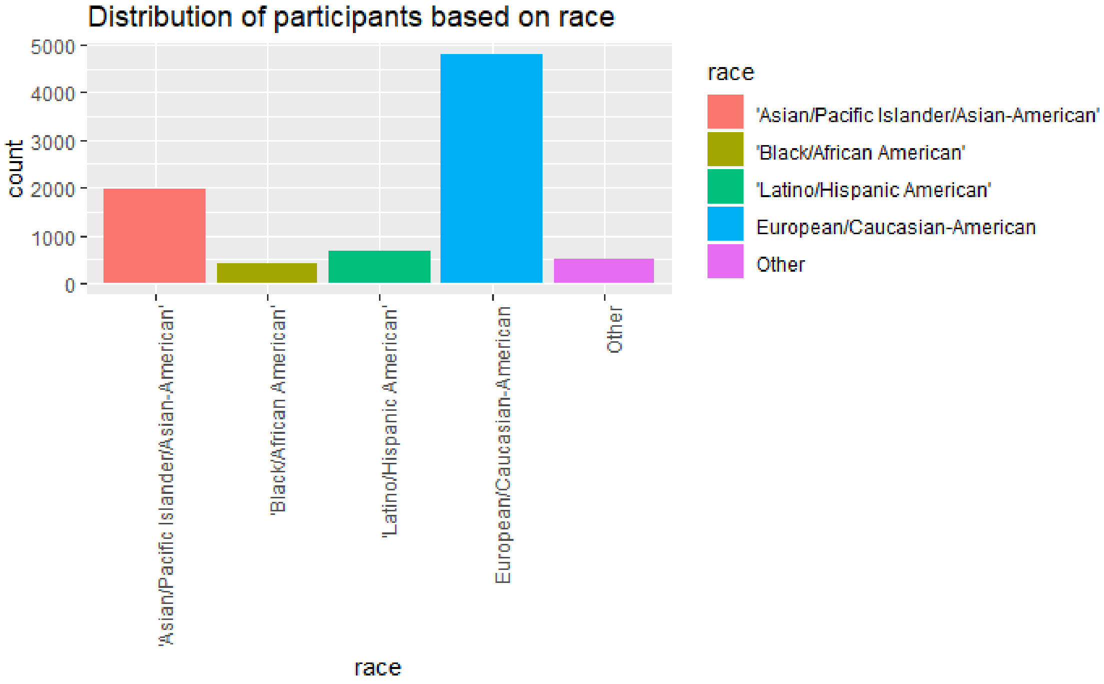
Let's look at the different categorical attributes one by one. First of all let's analyze the attribute **race**

```
## 'Asian/Pacific Islander/Asian-American'
##                                    1982
##                'Black/African American'
##                                     420
##              'Latino/Hispanic American'
##                                     664
##             European/Caucasian-American
##                                    4790
##                                   Other
##                                     522
```

As we can see, there are a total of 5 categories for race: "Asian/Pacific Islander/Asian-American" (which can also be referred to in short as Asian), "Black/African American" (Black in short), "Latino/Hispanic American" (Latino in short), "European/Caucasian-American" (White in short) and "Others".

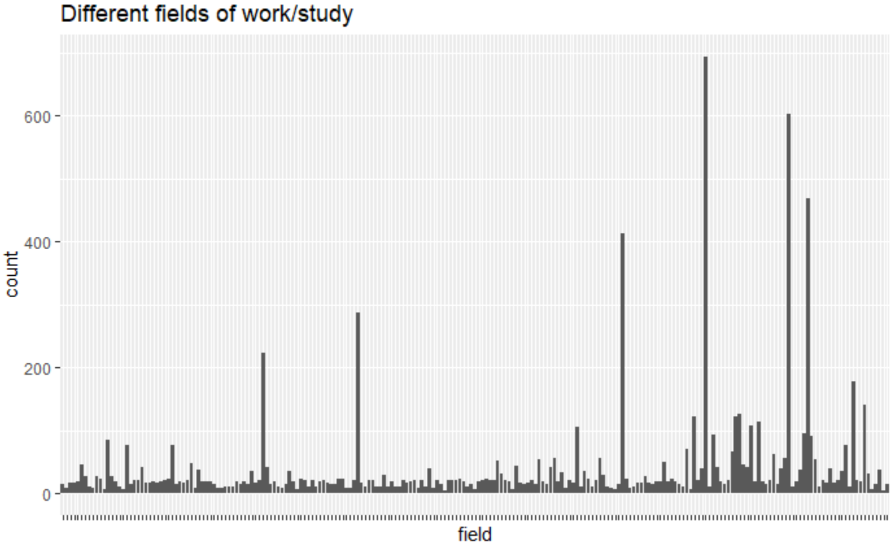
Most of the participants in the event are White (4790). The count for African America people is also significant with 1982 participants. People from other racial profiles are less in number compared to these two categories.

This information is summarised by the following plot

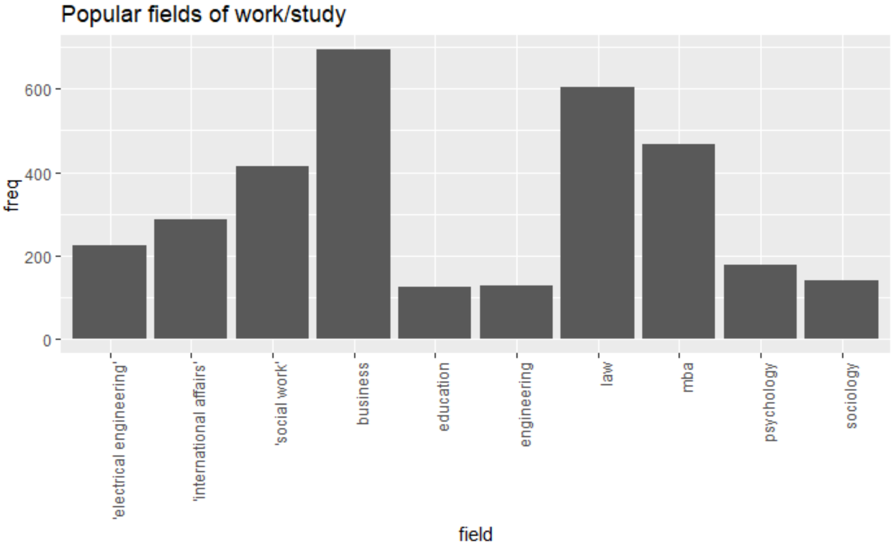


Occupational Distribution

Moving on to another categorical attribute: **field**



As we can see there are so many different fields where the participants are working in or studying. Let's summarise the top few categories with a plot:



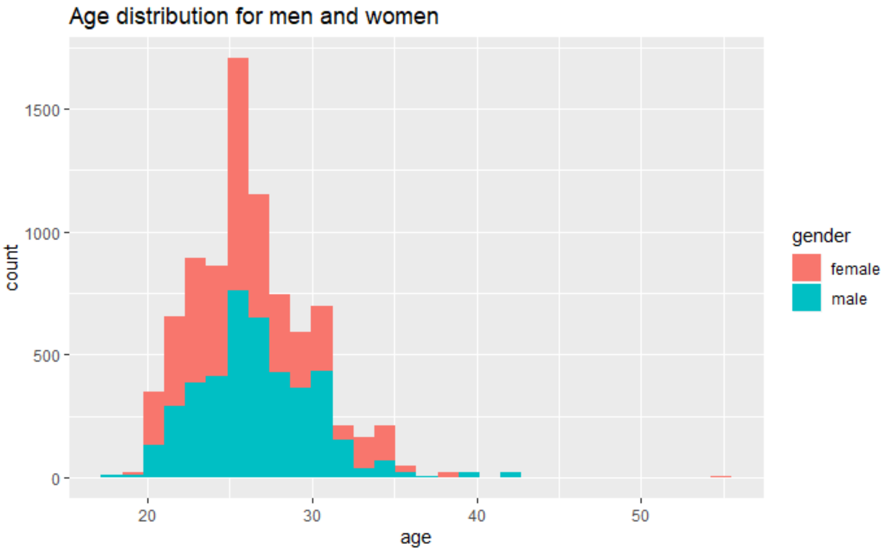
As seen from the plot, the most common field among the different participants is Business with 694 occurrences. It is closely followed by Law and MBA with frequencies of 604 and 468 respectively. But the important thing to note here is that no one field is dominating other fields in any significant way. The dataset seems to have a well diversified portfolio of people from different fields. This is a good feature to have in the dataset as it mimics the whole population much better.

Age Distribution

Now, let's look at the attribute **age**

Age is one of the most if not the most important attribute one looks for in their partner. As such it is very important to have a deeper look at the ages of all the people who participated in the event.

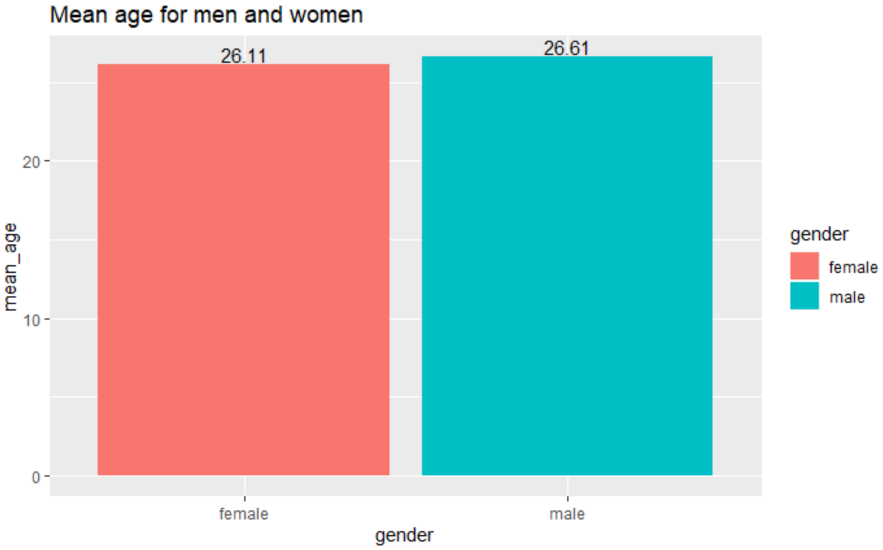
We will start by having a simple histogram for the age of the participants:



As we can see, the plot also distinguishes the age distribution for men and women. The spread of age is pretty much similar for both genders with the most number of participants from either gender being around the age of high 20s. There is also a significant amount of people with ages in the low twenties and low thirties. The graph also shows that not many people beyond the age of 35 showed up to participate in the event.

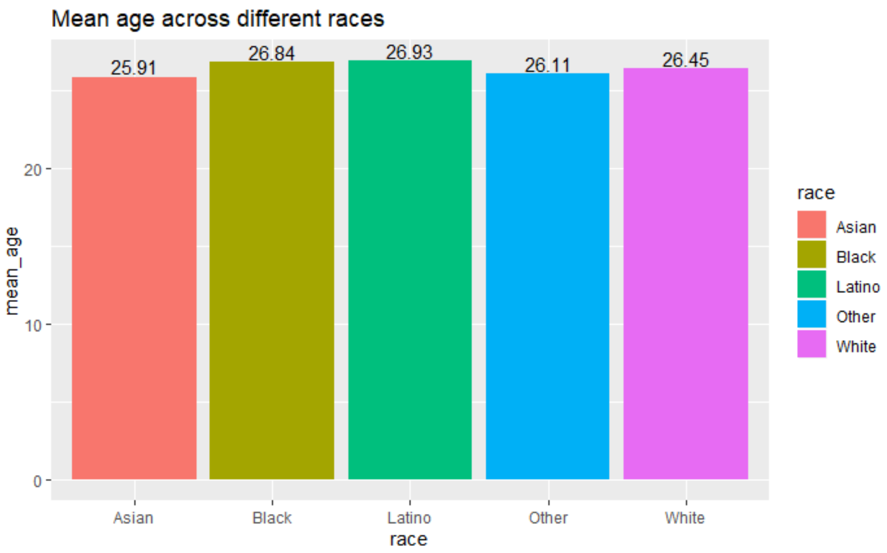
So it can be concluded from this plot that a very high percentage of people participating in this event are of the “right age” group in terms of eligibility to seek a romantic relationship.

This is also confirmed by the following plot where we look at the mean age for both genders



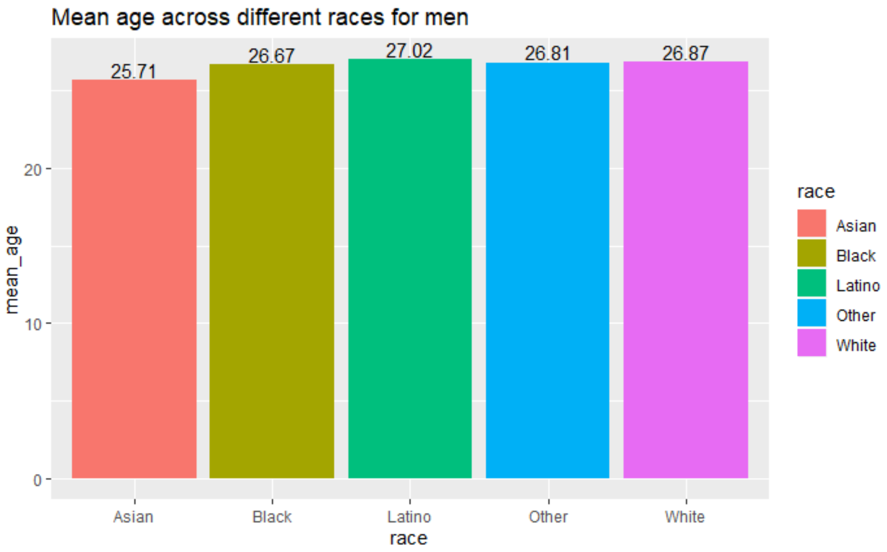
The plot shows that the average age for males and females participating in the event is almost exactly the same. The male participants are ever so slightly older than their female counterparts to be very accurate.

Another way of segregating participants in terms of their age is to look at the age distribution for participants from different racial profiles. This can be done by finding the mean age for participants with different races. The results are summarised by the following plot:



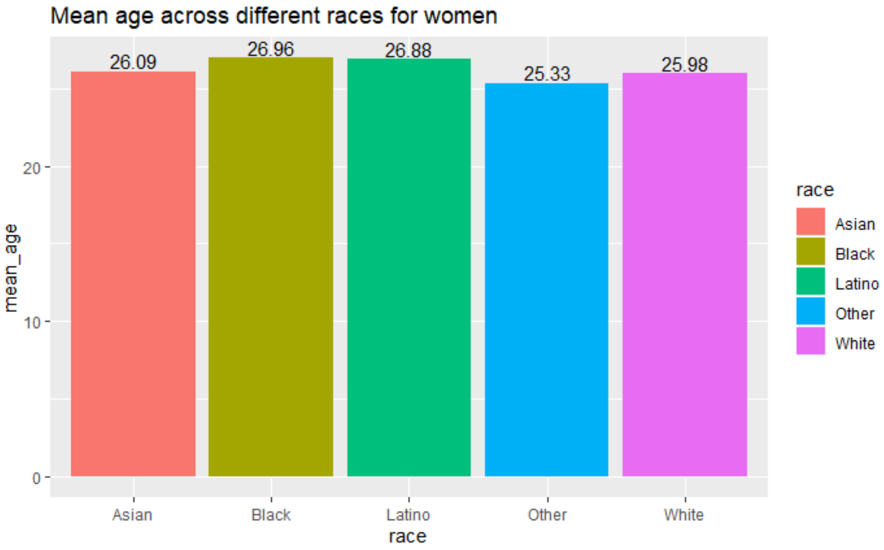
This plot also shows a very consistent distribution of age throughout the participants across different racial profiles. It is to be noted that out of all other races, people from Asian background have the lowest mean age of 25.91 while participant with Latin background have the highest mean age of 26.93. Between these two groups the difference is more than a year. Generally such a difference in age shouldn't mean much but it will be interesting to note the difference of ages of individuals in various dates. Participants that are either Black or White have almost the same age on average while people from Other categories are slightly younger than the rest except for the Asian group.

We can further analyze the age distribution of the participants of different races by separating the data for males and females as well. This information is summarised below:



The trend shown by this graph is almost similar to the previous graph. Asian men are still the youngest of the lot on average and Latino men are the oldest. The gap between the two is almost 1.5 years on average. But again, this difference shouldn't mean much in the grand scheme of things. White and Black men are again of comparable mean age but unlike in the previous graph, it is the Black men who are younger than the rest except Asian men among all other races. So among men, Black and Asian males are the youngest on average.

Looking at the similar plot for women:

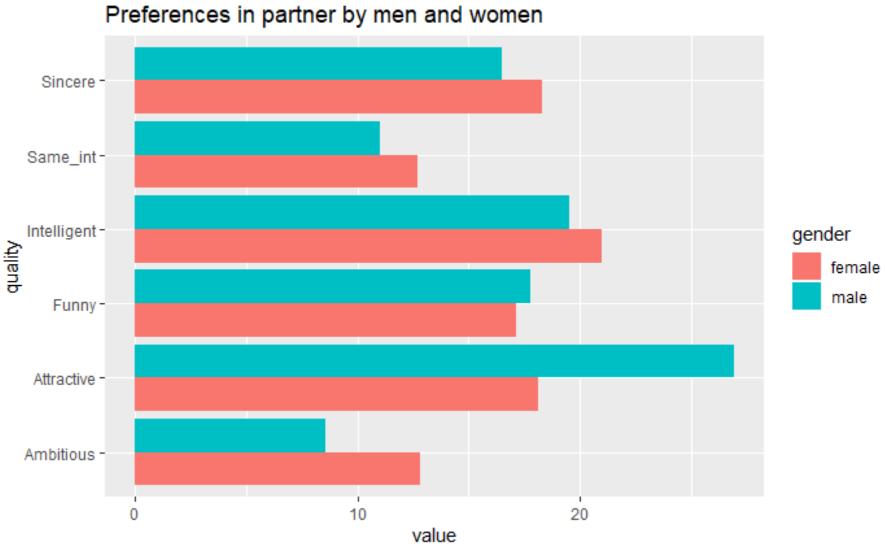


The distribution of age for females is different from the overall statistics. Unlike the previous analysis, Asian females are not the youngest group among other females. In Fact it is females from the category "Others" that take the cake in terms of being the youngest on average. The average of all categories is less than that of men as expected. White and Asian women are still relatively younger compared to Black and Latino women who are among the oldest on average among all women.

Analysis Related To Various Qualities

After analyzing these descriptive features of the dataset we can now look into more details of dataset in order to generate some insights into people's preferences, perception and expectations.

We will start by finding out what do different people look for in their partners. This analysis is done separately for both men and women in order to compare the preferences across genders. To do this analysis, we make use of the fields like **attractive_important** which basically depicts how important does a person rate attractiveness of their partner. This analysis is done for all different qualities that are available in the dataset. Since we want statistics on an aggregate for different genders, we suffice by taking the mean of these ratings as given by individuals.

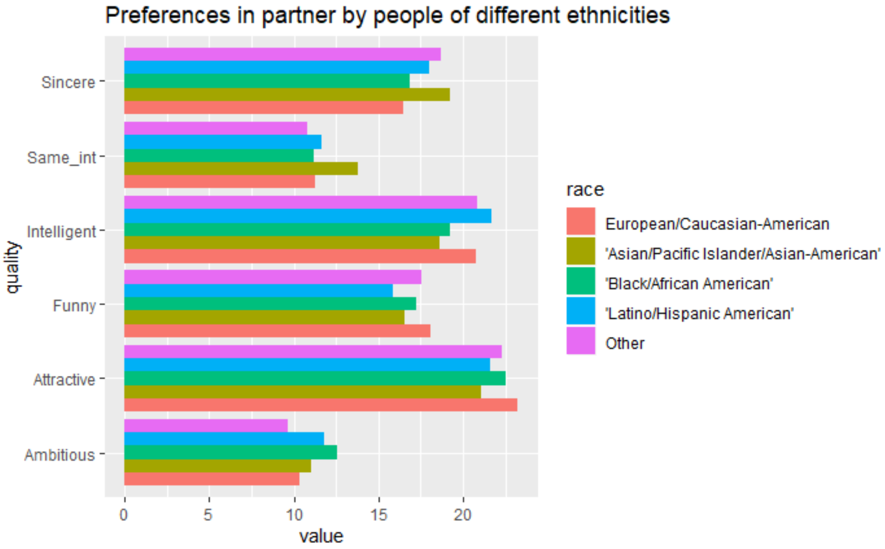


The plot above gives a lot of insights. Some of these revelations confirm common social beliefs but some denies them outright. For example, it is believed that men find it very important to have a partner who is attractive. This belief is completely supported by the data. Out of all other features, men give the most importance to their partner being attractive. And the difference between **attractiveness** and other qualities is quite significant in case of men. Women on the other hand do not give attractiveness as much importance as men do. In fact attractiveness is not even their most preferred quality in their partners. For women, the most important quality in their potential partner is **intelligence** but the gap between this and other qualities is not as significant. **This goes to show that women look for partners who are much more balanced across different qualities while men show an overwhelming inclination towards attractiveness.**

Looking at **sincerity**, both men and women give a decent amount of importance to this quality with women giving slightly more importance than men. Another quality which is significantly desired by both men and women is being **funny**. The importance for both men and women is quite close but men do show slightly more importance to it.

The qualities that are given relatively less importance by both men and women are **same interests** and **ambitions**. **Both men and women don't mind partners with somewhat different interests** but females do care about this a little more than males. Even though both men and women give less importance to ambition, **men really seem to care far less about their partners ambitions than women.**

We can do a similar analysis for people of different races.

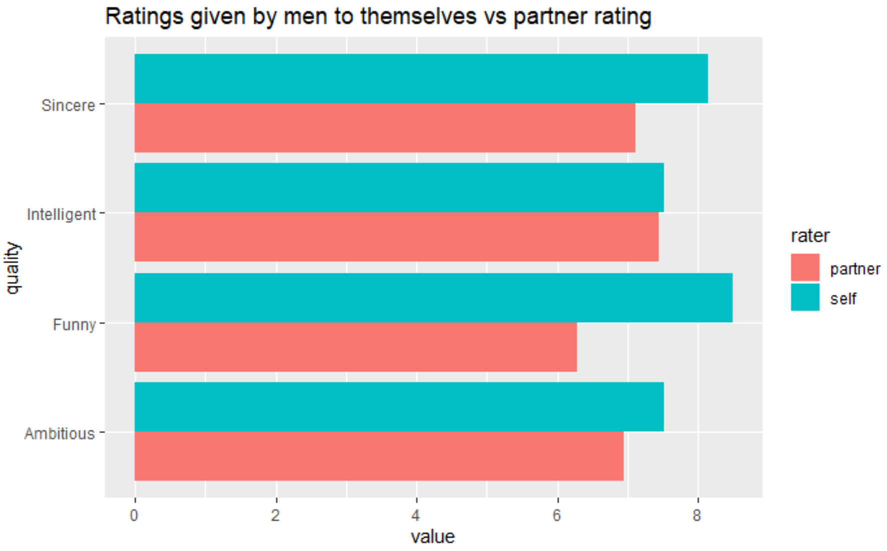


The plot shows some interesting results. Attractiveness has a clear advantage over other qualities unanimously for all groups of people with Caucasian participants showing the highest preference for an attractive partner. Some more interesting conclusions that can be drawn from the plot are:

1. People of Asian background give more importance to **sincerity** and same interests than people from other ethnicities.
2. European/Caucasian-American and Black/African-American people seem to care the least about **sincerity**.
3. Asian participants surprisingly care the least about **intelligence** among all other groups while Latin and White participants seem to rate intelligence the highest.
4. White and Black participants care more about their partner being **funny** than Latin and Asian participants.
5. **Ambition** is rated the highest by Black participants.

Moving on, we now analyse the observed differences between the ratings given by a person to oneself vs the ratings given to them by their potential partners. Like with previous cases, we do this analysis separately for both genders.

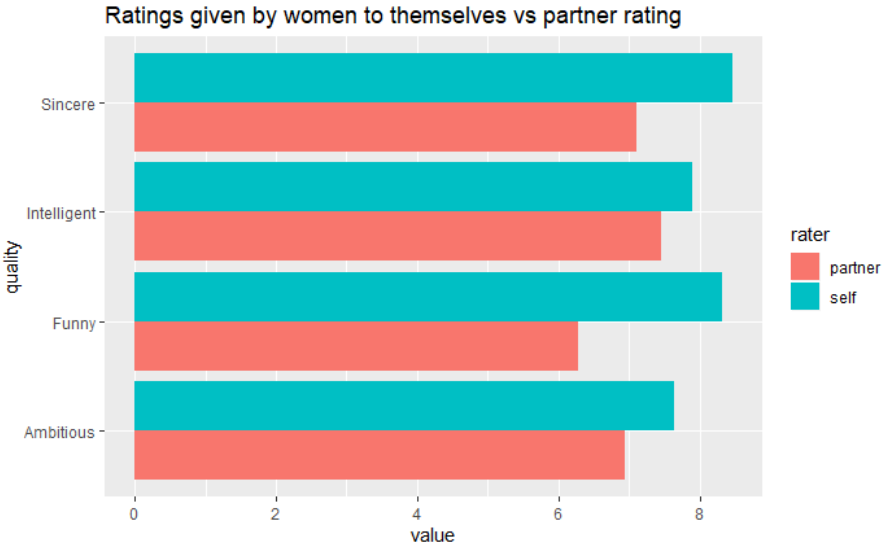
First, let us look at the how **men** perceive themselves across different attributes vs how they were rated by their partners.



Right off the bat it is clear that there is overestimation of oneself by men across all qualities. Some interesting conclusions from the plot are:

1. Men rate their **intelligence** nearly the same as women opposite to them do.
2. The biggest misconception men have about themselves is being more **funny** than what their potential partners think.
3. The amount of overestimation men do for qualities like **sincerity**, **attractiveness** and **ambition** is about the same when compared to the ratings given by women.

Moving on to the same analysis for women



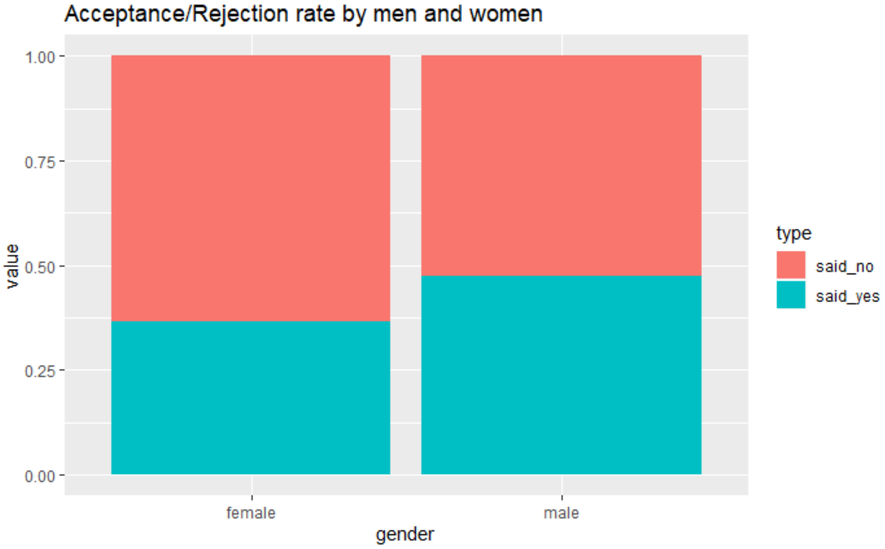
A similar trend is seen in case of women as well. They also seem to overestimate their attributes in all aspects when compared to the partners' ratings. Some other conclusions from the plots are:

1. The gap between the self perception of being **funny** and what the partner rates them to be is also the largest in case of women.

2. There is also a significant gap between the ratings given to self by women on **sincerity** and **ambition** vs the ratings given on the same quality by men.
3. The difference between the perceived and rated values of **attractiveness** and **intelligence** is less for women.

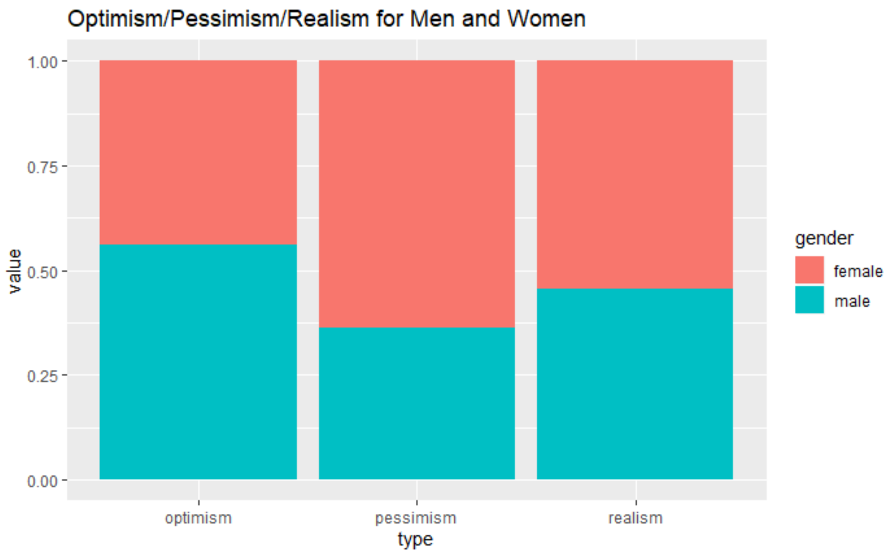
Rate of Acceptance/ Optimism/ Pessimism

Next we will look into the acceptance rate across different genders and ethnicities. By acceptance rate we generally refer to the fraction of people belonging to a certain group that actually say yes to their partners. This analysis does not take into account other factors, rather it just serves as a metric to define which groups are seemingly less strict when saying yes to potential romantic relationships.



It is quite clear from the plot that **Men say yes to potential partners more often than women and consequently women say no more than men.**

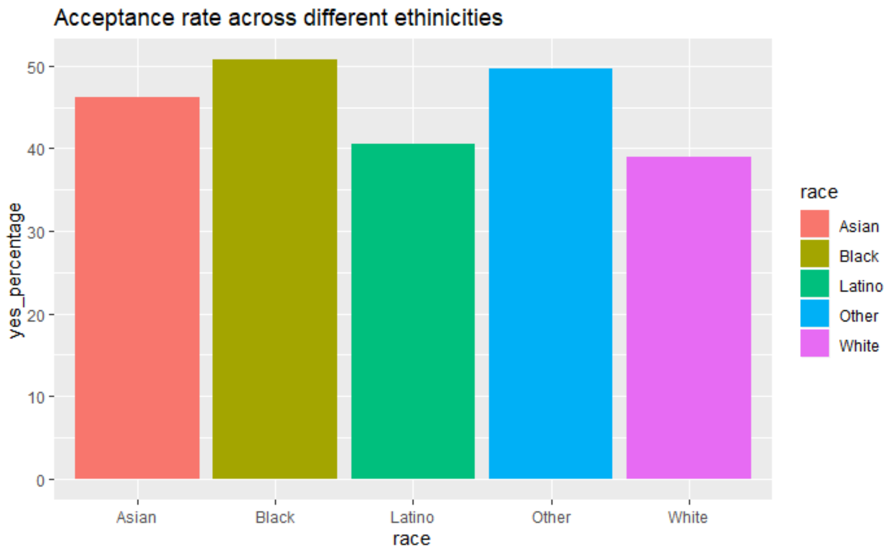
Next we look at fractions when a person thinks the opposite person is into them while the partner says no (we term this as optimism) and the times when a person does not think the partner will like them but in reality the partner did say yes to them (we term this pessimism). Along the same terms, we define realism as when a person correctly anticipates the partners liking/disliking towards them.



The following can be concluded from the plot:

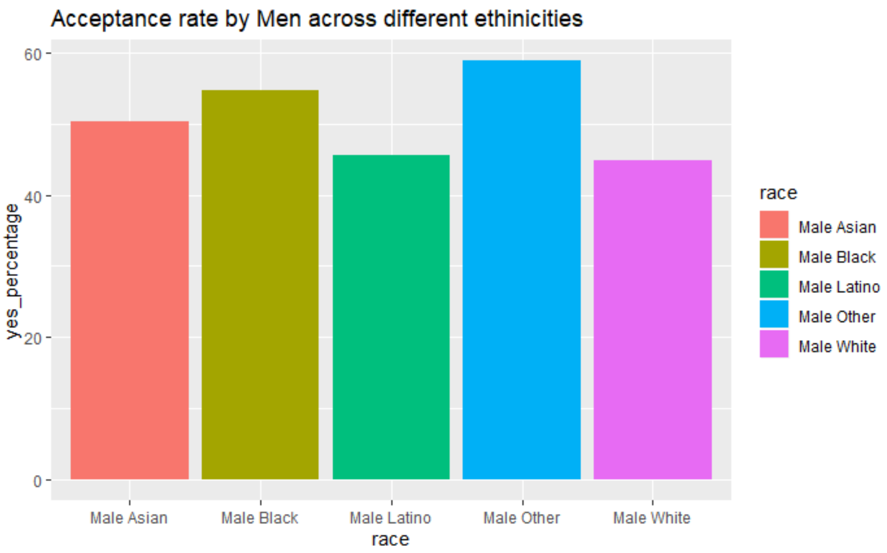
1. More men are optimistic than women and the difference is also significant.
2. More women are pessimistic than men and the difference is large in this case too between the two genders.
3. The above observations conclude what the plot also shows that **men are generally more optimistic about their chances and women are pessimistic in this regard.** Also women seem to have a slightly better idea of judging whether their partner is interested in them or not (realism).

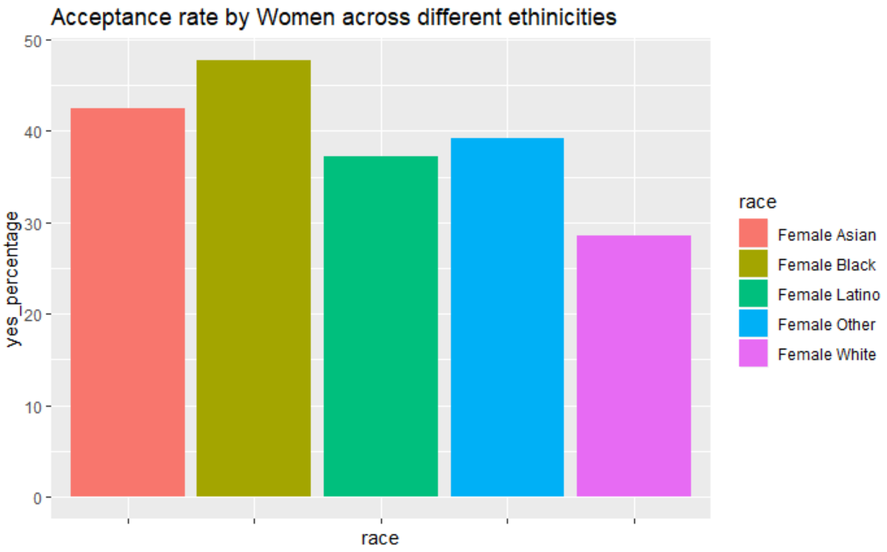
Now, let's look at percentages of various ethnic groups that say yes to their partners.



The graph shows that **Black participants say yes more than any other ethinic groups while White participants say yes the least.**

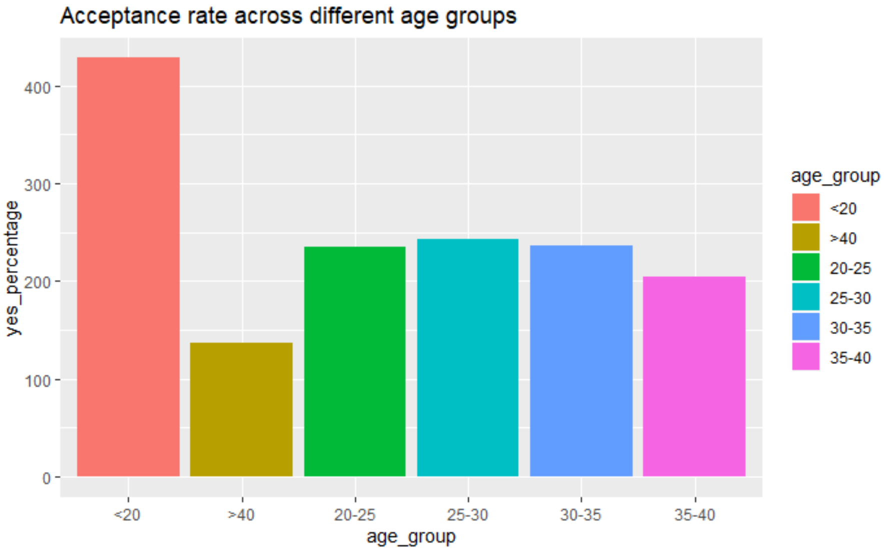
This can also be decomposed for both genders:


Acceptance rate by Men across different ethinicities


Acceptance rate by Women across different ethinicities

The above two plots simply corroborates the overall results.

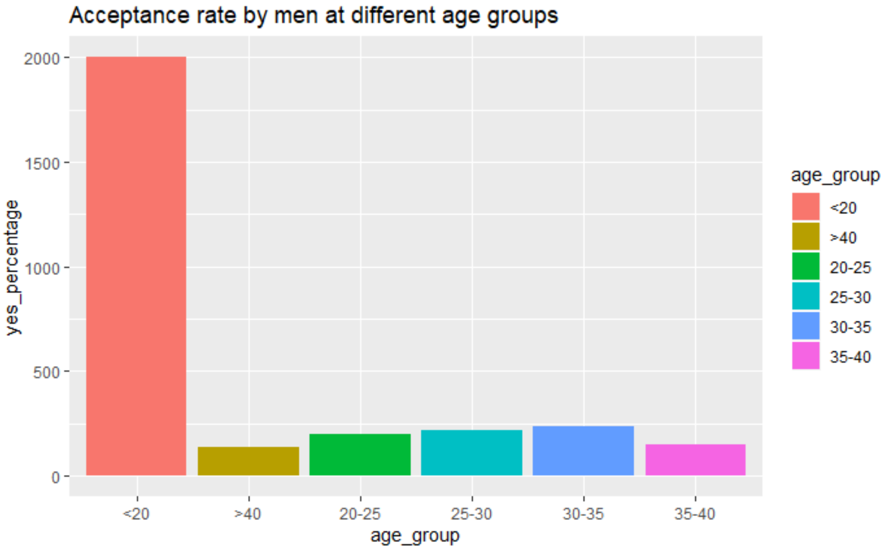
Next we look at acceptance rate for people in different age groups.

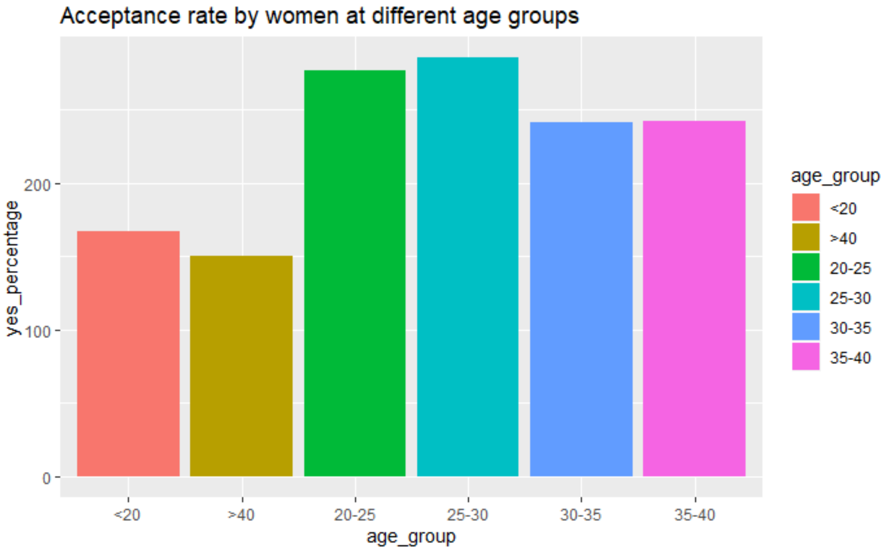


The following conclusions can be drawn from this plot:

1. Participants below the age of 20 say yes a lot more than any other age group. The difference is almost a 100 percent. We can say that **very young people are very generous in their acceptance for potential romantic partners.**
2. People above the age of 40 say yes to potential partners the least. We can say that **older people are very strict in their selection process.**

Doing the same analysis for both men and women separately.



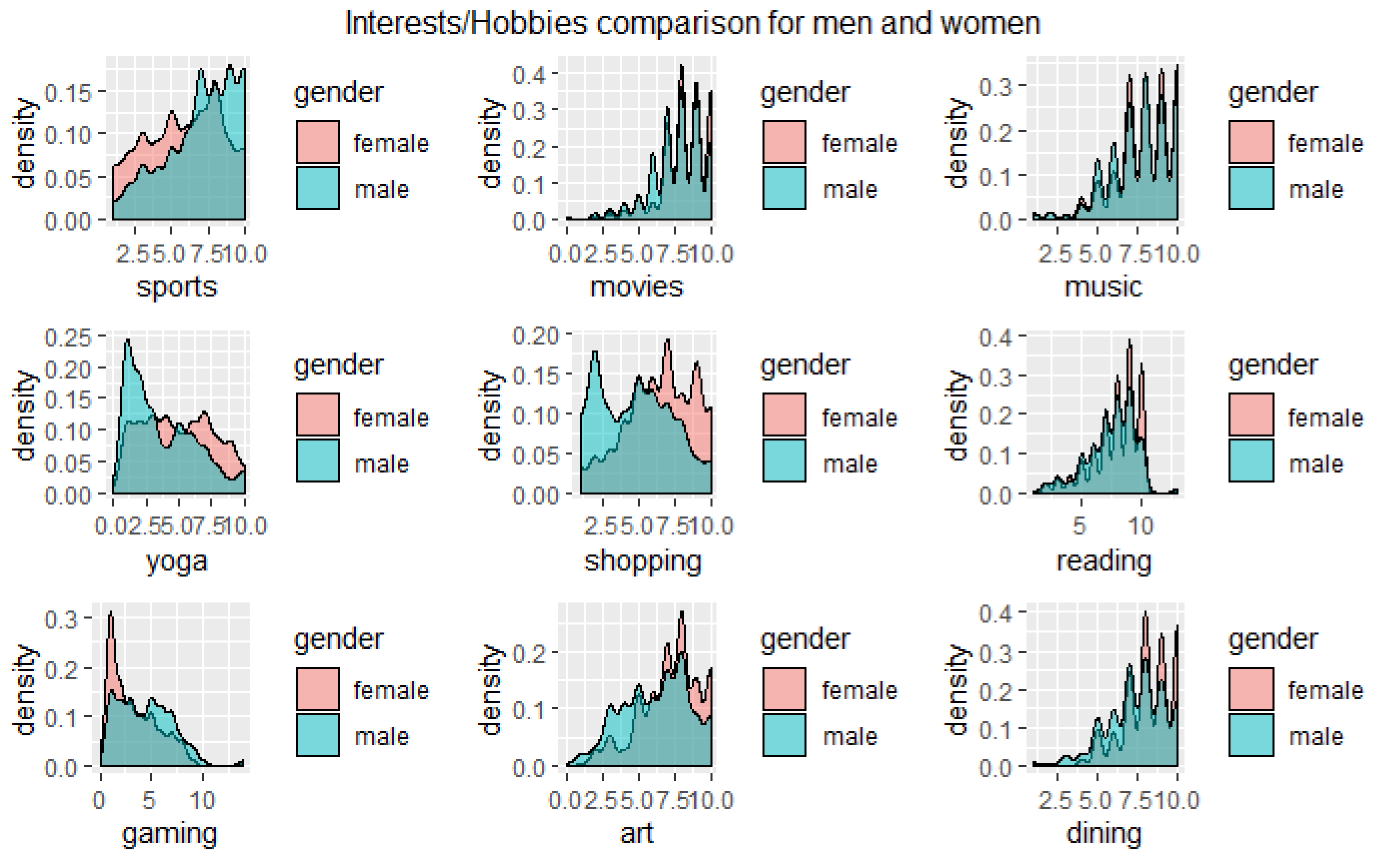


The results from plots for men and women are quite different. Following useful conclusions can be drawn from these:

1. The percentage of men below the age of 20 that say yes to others is extraordinarily large compared to all other groups. We can say that **young men are extremely eager to say yer to various partners.**
2. The percentage of women below the age of 20 and above the age of 40, that say yes to partners is lower than all other age groups among women. We can say that **women between the ages of 20 and 40 are more likely to say yes than women that are too young or old.**

Hobbies and Interests Analysis

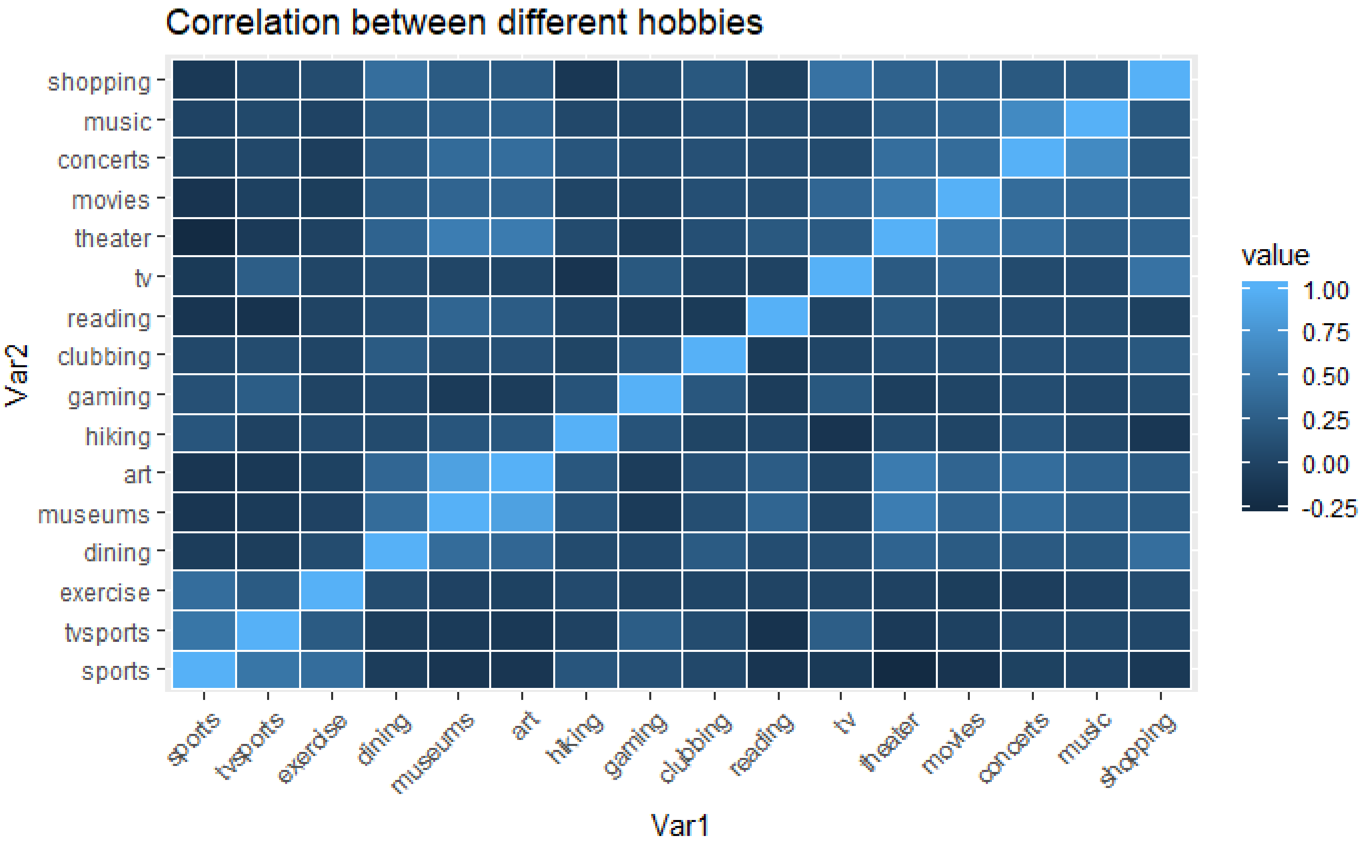
The dataset also contains a lot of information regarding various hobbies/interests each person has. This has some potential to analyse the differences between men and women in terms of their various interests and hobbies. The following plots summarise some of these hobbies:



Looking at the various plots in the grid above we can observe the following:

1. Men show a higher propensity towards **Sports** and **Yoga** although difference isn't as much as one would expect.
2. Women surprisingly show more interest in **gaming** then men.
3. Women have a higher propensity towards **arts**, **shopping** and **dining**.
4. Activities like **movies**, **music**, and **reading** are liked almost equally by men and women.

We can further look at the correlation between different hobbies:



In the above plot, the lighter color tiles represent a higher correlation. As such we can see that the results are quite obvious for this dataset. Museums and arts seem to have a very high correlation meaning that any person who likes arts also like museums. Similarly, concerts and music have a higher correlation. Observing the dark tiles, we can see that there is very low correlation between sports and arts/museums.

Similar plots made separately for men and women show similar results.

Methodology

In this section we will look at different methods for performing predictive analysis on the data.

Predicting Overall Match

First task is to take all relevant information available for both participants involved in each particular "date" and fit a model for their match. Before doing this analysis, we need to remove certain features from the dataset. Features like **decision** and **decision_o** basically decide the final outcome, hence these cannot be used in this analysis. We will perform a predictive analysis of these two features in the subsequent sections.

The first step is to split the dataset into three sets for **training**, **validation** and **testing**. Out of a total of **8378** rows in the dataset, we use **1500** for validation, another **1500** for testing and the rest are used for training.

Now, we try different methods on our dataset. We also introduce each method briefly for the sake of completion.

QDA

QDA[4] is not really that much different from LDA except that you assume that the covariance matrix can be different for each class and so, we will estimate the covariance matrix $\Sigma_k \Sigma_k$ separately for each class k, k =1, 2, . . . , K.

Quadratic discriminant function:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log \Pi_k$$

This quadratic discriminant function is very much like the linear discriminant function except that because Σ_k, the covariance matrix, is not identical, you cannot throw away the quadratic terms. This discriminant function is a quadratic function and will contain second order terms.

Classification rule:

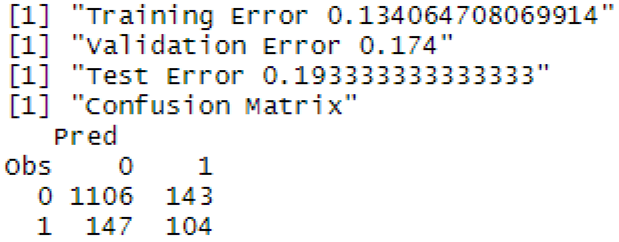
$$G(x) = argmax_k \delta_k(x)$$

The classification rule is similar as well. You just find the class k which maximizes the quadratic discriminant function.

The decision boundaries are quadratic equations in x.

QDA, because it allows for more flexibility for the covariance matrix, tends to fit the data better than LDA, but then it has more parameters to estimate. The number of parameters increases significantly with QDA. Because, with QDA, you will have a separate covariance matrix for every class. If you have many classes and not so many sample points, this can be a problem.

We ran the QDA model fit on the given data obtained the following results:



```
[1] "Training Error 0.134064708069914"
[1] "Validation Error 0.174"
[1] "Test Error 0.193333333333333"
[1] "Confusion Matrix"
   Pred
Obs    0    1
  0 1106  143
  1  147  104
```

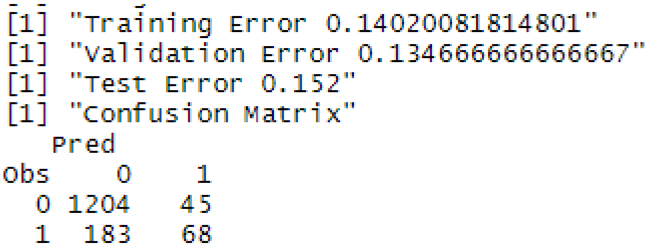
Logistic Regression

Logistic Regression[4] is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic Regression is one of the most popular ways to fit models for categorical data, especially for binary response data in Data Modeling. It is the most important (and probably most used) member of a class of models called generalized linear models. Unlike linear regression, logistic regression can directly predict probabilities (values that are restricted to the (0,1) interval); furthermore, those probabilities are well-calibrated when compared to the probabilities predicted by some other classifiers, such as Naive Bayes. Logistic regression preserves the marginal probabilities of the training data. The coefficients of the model also provide some hint of the relative importance of each input variable.

Logistic Regression is used when the dependent variable (target) is categorical. Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequences in real time.

We ran logistic regression on the data set and here are the results:



```
[1] "Training Error 0.14020081814801"
[1] "Validation Error 0.134666666666667"
[1] "Test Error 0.152"
[1] "Confusion Matrix"
   Pred
Obs    0    1
  0 1204   45
  1  183   68
```

Naive Bayes

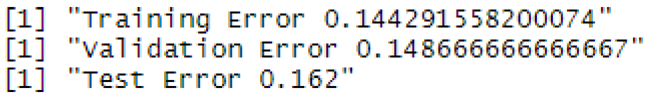
Naive Bayes[4] is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

We ran the Naive Bayesian mechanism on the data set and here are the results:



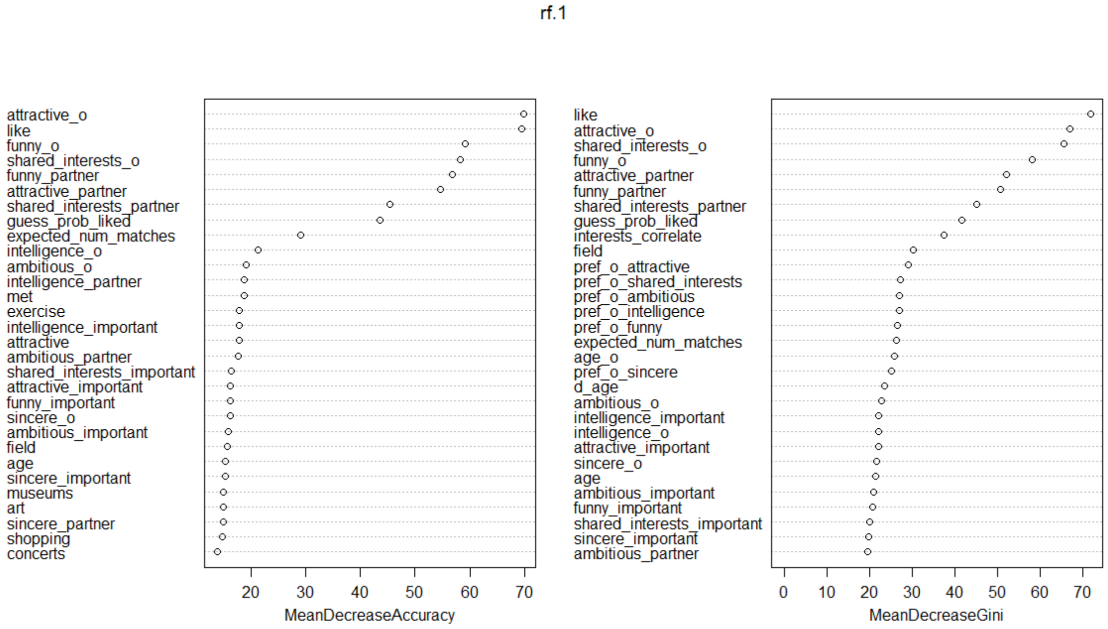
```
[1] "Training Error 0.144291558200074"
[1] "Validation Error 0.148666666666667"
[1] "Test Error 0.162"
```

Random Forest

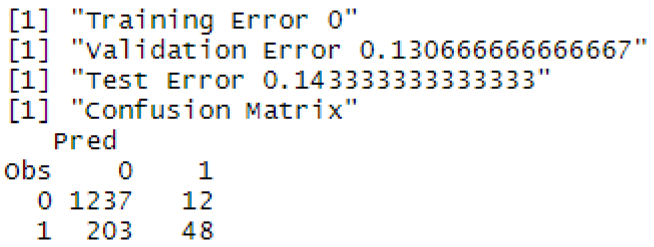
A Random Forest[4] consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. Alternatively, for regression problems, the tree response is an estimate of the dependent variable given the predictors. The Random Forest algorithm was developed by Breiman.

A Random Forest consists of an arbitrary number of simple trees, which are used to determine the final outcome. For classification problems, the ensemble of simple trees vote for the most popular class. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy.

Using RandomForest method we get the following plot depicting the importance of different attributes in the model



After running the model on the dataset, we get the following results:



```
[1] "Training Error 0"
[1] "Validation Error 0.130666666666667"
[1] "Test Error 0.143333333333333"
[1] "Confusion Matrix"
   Pred
Obs    0    1
  0 1237   12
  1  203   48
```

SVM

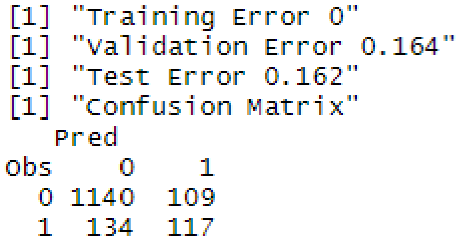
Support Vector Machine (SVM)[4] is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables a dummy variable is created with case values as either a 0 or 1. Thus, a categorical dependent variable consisting of three levels, say (A, B, C), is represented by a set of three dummy variables:

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

To construct an optimal hyperplane, SVM employs an iterative training algorithm, which is used to minimize an error function.

In SVM, we did tune the model for various hyperparameters like **gamma** and **cost** and used the parameters that gave the best results during the cross validation tuning process. The tuned values are **gamma = 0.01** and **cost = 100**.

We ran SVM on the given data set and here are the results:



```
[1] "Training Error 0"
[1] "Validation Error 0.164"
[1] "Test Error 0.162"
[1] "Confusion Matrix"
   Pred
Obs    0    1
  0 1140  109
  1  134  117
```

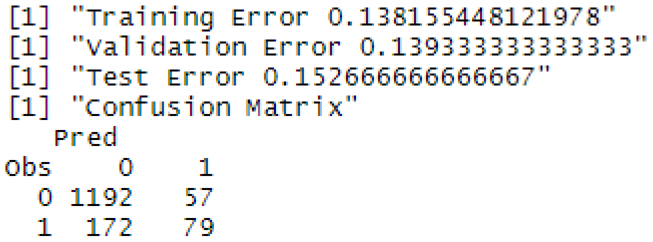
Neural Networks

Neural networks[4] have seen an explosion of interest over the last few years, and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as finance, medicine, engineering, geology and physics. Indeed, anywhere that there are problems of prediction, classification or control, neural networks are being introduced. Neural networks are also intuitively appealing, based as they are on a crude low-level model of biological neural systems.

Neural networks are very sophisticated modeling techniques capable of modeling extremely complex functions. In particular, neural networks are nonlinear (a term which is discussed in more detail later in this section). For many years linear modeling has been the commonly used technique in most modeling domains since linear models have well-known optimization strategies. Where the linear approximation was not valid (which was frequently the case) the models suffered accordingly. Neural networks also keep in check the curse of dimensionality problem that bedevils attempts to model nonlinear functions with large numbers of variables.

Neural networks learn by example. The neural network user gathers representative data, and then invokes training algorithms to automatically learn the structure of the data. Although the user does need to have some heuristic knowledge of how to select and prepare data, how to select an appropriate neural network, and how to interpret the results, the level of user knowledge needed to successfully apply neural networks is much lower than would be the case using (for example) some more traditional nonlinear statistical methods.

We ran the neural net on the given data set and here are the results:



```
[1] "Training Error 0.138155448121978"
[1] "Validation Error 0.139333333333333"
[1] "Test Error 0.152666666666667"
[1] "Confusion Matrix"
   Pred
Obs    0    1
  0 1192   57
  1  172   79
```

XGBoost

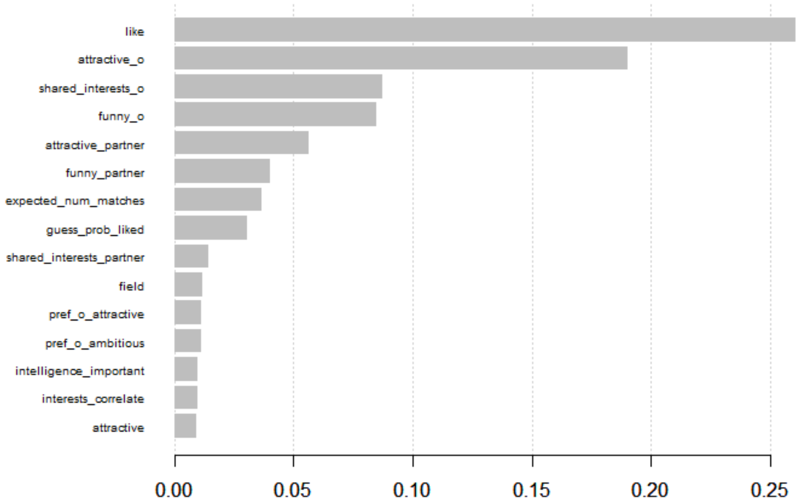
Gradient boosting[4] is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Gradient boosting involves three elements: a loss function to be optimized, a weak learner to make predictions and an additive model to add weak learners to minimize the loss function.

The loss function used depends on the type of problem being solved. It must be differentiable, but many standard loss functions are supported and you can define your own. For example, regression may use a squared error and classification may use logarithmic loss. A benefit of the gradient boosting framework is that a new boosting algorithm does not have to be derived for each loss function that may want to be used, instead, it is a generic enough framework that any differentiable loss function can be used.

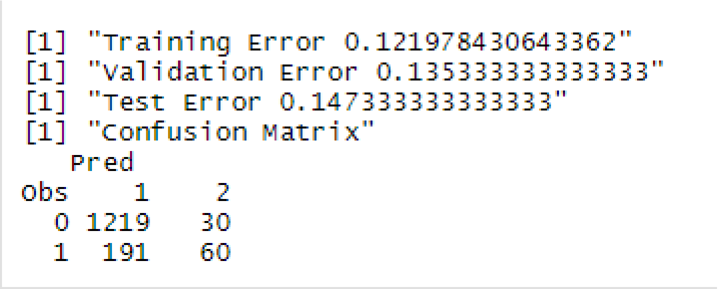
Decision trees are used as the weak learner in gradient boosting. Specifically regression trees are used that output real values for splits and whose output can be added together, allowing subsequent models outputs to be added and “correct” the residuals in the predictions. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss. Initially, such as in the case of AdaBoost, very short decision trees were used that only had a single split, called a decision stump. Larger trees can be used generally with 4-to-8 levels. It is common to constrain the weak learners in specific ways, such as a maximum number of layers, nodes, splits or leaf nodes. This is to ensure that learners remain weak, but can still be constructed in a greedy manner.

Trees are added one at a time, and existing trees in the model are not changed. A gradient descent procedure is used to minimize the loss when adding trees. Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error. Instead of parameters, we have weak learner sub-models or more specifically decision trees. After calculating the loss, to perform the gradient descent procedure, we must add a tree to the model that reduces the loss (i.e. follow the gradient). We do this by parameterizing the tree, then modify the parameters of the tree and move in the right direction by (reducing the residual loss. Generally this approach is called functional gradient descent or gradient descent with functions. The output for the new tree is then added to the output of the existing sequence of trees in an effort to correct or improve the final output of the model. A fixed number of trees are added or training stops once loss reaches an acceptable level or no longer improves on an external validation dataset.

Using the gradient boost method, we ran the variable importance on the data set. The following plot shows the importance of the different attributes as per the gradient boost method:



After running the model fit on the data set, here are the results:



```
[1] "Training Error 0.121978430643362"
[1] "Validation Error 0.135333333333333"
[1] "Test Error 0.147333333333333"
[1] "Confusion Matrix"
   Pred
Obs    1    2
  0 1219   30
  1  191   60
```

K Nearest Neighbours

In pattern recognition, the k-nearest neighbors algorithm (k-NN)[4] is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

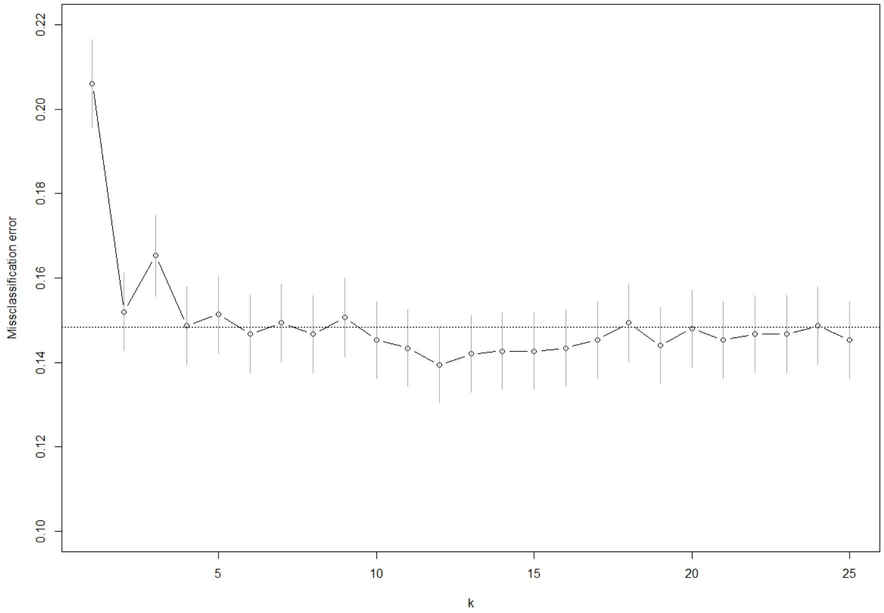
In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.[2]

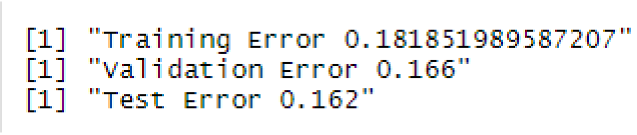
The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

Using different values of k (number of neighbours) we plot the error rate:



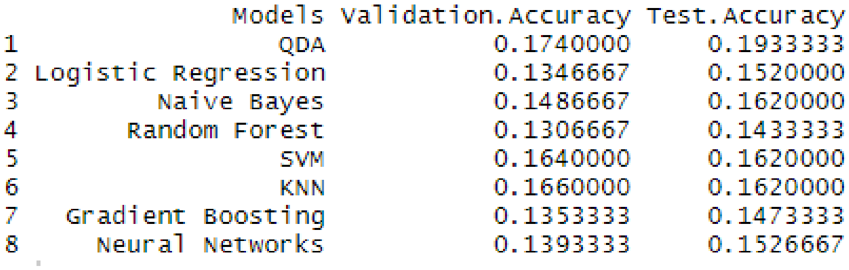
As shown by the plot, **k=12** shows the lower error, so we will use that to fit the model. Here are the results after the model fit:



```
[1] "Training Error 0.181851989587207"
[1] "Validation Error 0.166"
[1] "Test Error 0.162"
```

Comparison of different model performances

The performance of various models on the data sets are summarized below:



```
               Models Validation.Accuracy Test.Accuracy
1                 QDA           0.1740000     0.1933333
2 Logistic Regression           0.1346667     0.1520000
3         Naive Bayes           0.1486667     0.1620000
4       Random Forest           0.1306667     0.1433333
5                 SVM           0.1640000     0.1620000
6                 KNN           0.1660000     0.1620000
7   Gradient Boosting           0.1353333     0.1473333
8     Neural Networks           0.1393333     0.1526667
```

Predicting Individual Decisions

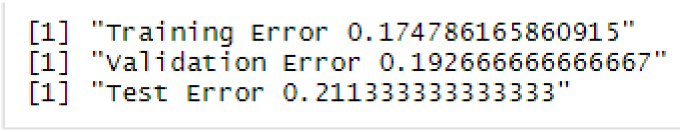
The second thing we want to achieve is that given all the attributes of a person and their potential partner, we can predict the individual's decision. This analysis is important as it will reflect how individuals make decisions based on various factors.

The divide into training, validation and test sets remains the same for this task as well. But we need to remove attributes like individual decisions and overall match in order to perform this task.

We decided to use methods **Gradient Boosting** and **K nearest neighbours** as they performed pretty well in the previous tasks.

XGBoost

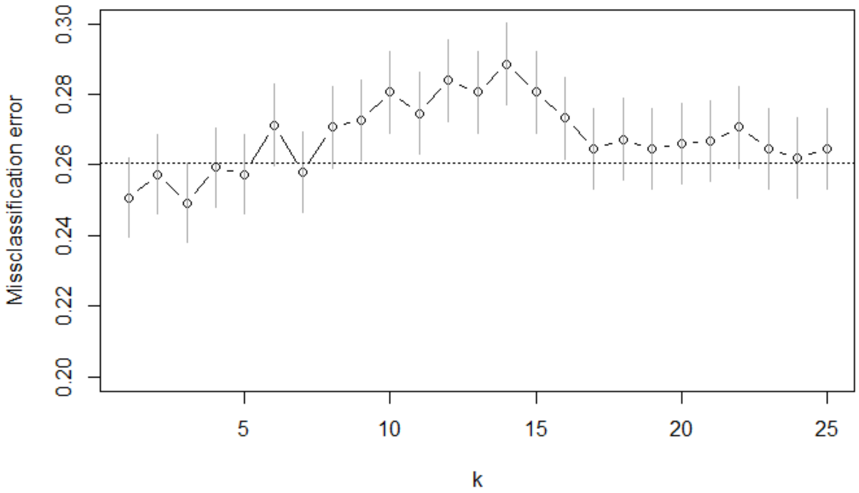
After running XGBoost as done in the previous section but with appropriately prepared data set we get the following results:



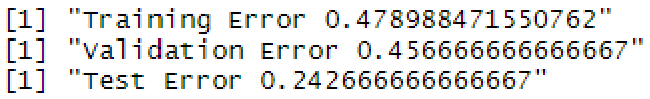
```
[1] "Training Error 0.174786165860915"
[1] "Validation Error 0.192666666666667"
[1] "Test Error 0.211333333333333"
```

K Nearest Neighbours

We apply KNN in the same manner as in the previous section. After cross validating with different values of k, we get the following plot showing the results of the cross validation:



As seen by the plot the value of k with least error is 3. We use this value and here are the results:



```
[1] "Training Error 0.478988471550762"
[1] "Validation Error 0.456666666666667"
[1] "Test Error 0.242666666666667"
```

Surprisingly, the results for this task aren't as accurate as the previous task. **It is harder to predict individual decisions than it is to predict overall match.**

Gender Differences

In this section, we look to explore some of the gender differences among the participants of the event. In particular, we look to find how do participants of different genders rate different qualities in their partners.

For this analysis, we only keep six attributes related to ratings for qualities like attractiveness, sincerity, intelligence, etc. We also make use of the overall decision attribute to draw some conclusions. And finally we split the data into two sets, one for men and another for women.

Here we use **RandomForest** method as it is good for indicating the importance of various attributes in a model and it also yielded good results in the previous sections.

The results are as follows:

As we can see from the plot, male participants give more importance to attractiveness than their female counterparts. While attractiveness and being funny are the two most desirable categories as per the predictive analysis, itelligence, sincerity and ambition are least desirable for both genders.

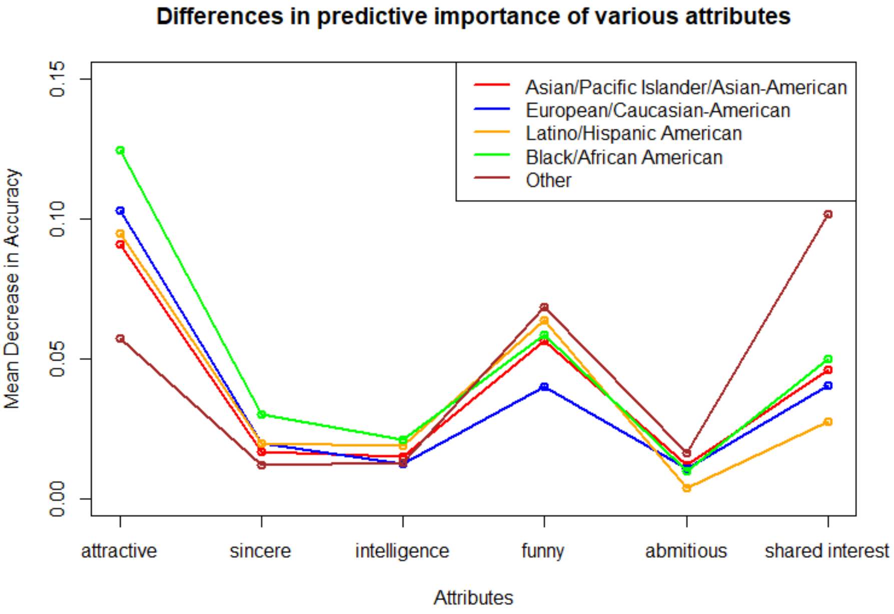
Ethnic Differences

In this section, we look to explore some of the differences based on ethnic groups of the participants of the event. In particular, we look to find how do participants of different races rate different qualities in their partners.

For this analysis, we only keep six attributes related to ratings for qualities like attractiveness, sincerity, intelligence, etc. We also make use of the overall decision attribute to draw some conclusions. And finally we split the date into different sets each for different ethnic groups.

Here we use **RandomForest** method as it is good for indicating the importance of various attributes in a model and it also yielded good results in the previous sections.

The results are as follows:

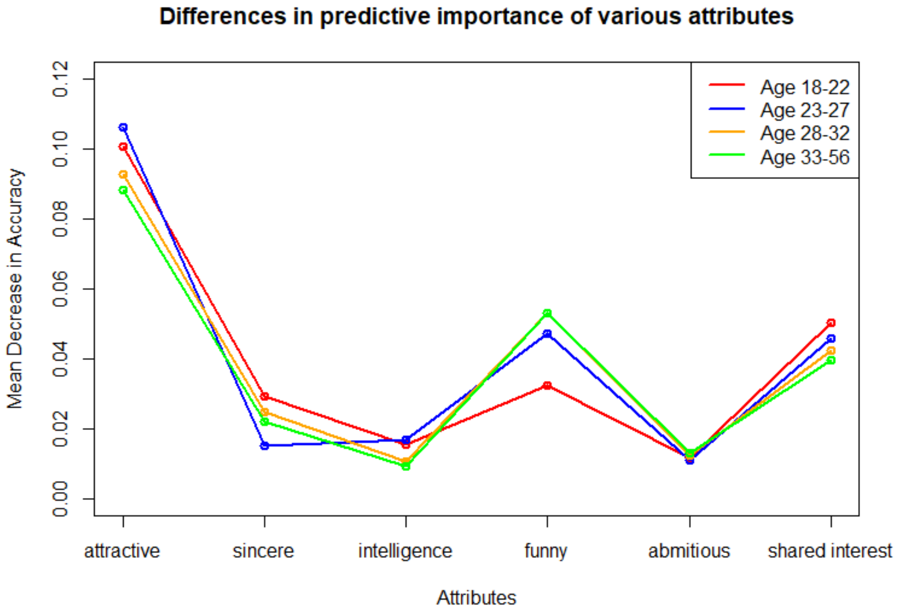


As expected, attractiveness leads the charts here as well. White and Black participants give more importance of attractiveness than other groups. Whereas Asian and Latino participants show more importance than other groups to their partner being funny. Shared interest is rated the highest by Asian and Black participants as well. Latin participants show the least inclination towards an ambitious partner out of all other ethnic groups.

Age Differences

In this section, we look to explore some of the differences based on varying age groups of the participants in the event. In particular, we look to find how do participants of different ages rate different qualities in their partners.

For this analysis, we only keep six attributes related to ratings for qualities like attractiveness, sincerity, intelligence, etc. We also make use of the overall decision attribute to draw some conclusions. And finally we split the date into different sets each for different age group.



Again, in case of different age groups as well, it is attractiveness that seems to be the most effective as per the predictive analysis. And this inclination is very close in all age groups, although participants of age less than 23 show slightly higher inclination. Younger participants (age 18) show less importance to their partner being funny while they care the most among all other age groups about their partner having the same interests as they do.

Conclusions

Based on the questions we listed out in the beginning of this report, here are our conclusions:

1. Given a set of attributes pertaining to potential partners, we are able to predict whether the two people are ready to date each other with XX% accuracy using ABC method.
2. Based on a person's own preferences, and the attributes of the potential partner, we are able to predict whether the person will be interested in the opposite person using XYZ methods. The accuracy of these predictions is lower.
3. Predictive analysis reveals that the most desirable attributes for both Men and Women are attractiveness and funny. Men also give more importance to attractiveness than women. The importance of all other qualities for men and women is similar.
4. Looking at the preferences across participants from different racial groups, predictive analysis reveals that Black and White participants give more weightage to attractiveness than other groups. Latino and Asian participants tend to give more importance than other groups to their partner being funny.
5. People in different age groups still show similar trends in terms of their preferred qualities as per the predictive analysis. Younger people tend to show more importance for attractiveness and shared interests, whereas people who are older tend to give more importance to than young people to their partner being funny.

Future Work

In our conclusions, we state that we could more accurately predict a match than an individual decision. This is counter-intuitive and can be further studied to determine the reason for this. The dataset contains a lot of features and a lot more analysis can be performed. One such thing can be to see the differences in factors that influence partner selection for people from various fields of work. Another interesting future work may include finding inherent clusters among the people on the basis of various features available.

The best accuracy we could achieve for predicting a match was less than 90%. This may signify that the data quality can be much better and many more features could have been included in the dataset. Some of our suggestions are - importance of race for partner, religion, and interest ratings of partner, would the person want someone from their own field of work. Some other concerns we had with the data was the missing data. It is inevitable that we have missing data when collected from real world, but imputing the data generates some bias.

Contributions

Puneeth's contributions include: Data Collection/Selection, preprocessing of data, cleaning data, predictive analysis, etc.

Govind's contributions include: Data Collection/Selection, descriptive analysis, report writing, predictive analysis, etc.

Appendix

Literature

Two research papers have been published which specifically analyses the speed dating dataset.

The first paper[1] titled “Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment” studies the differences in various factors that influences men and women to choose a partner. They find that women put greater weight on the intelligence and the race of partner, while men respond more to physical attractiveness. Moreover, men do not value women’s intelligence or ambition when it exceeds their own. Also, they find that women exhibit a preference for men who grew up in affluent neighborhoods. Male selectivity is invariant to group size, while female selectivity is strongly increasing in group size.

“Racial Preferences in Dating: Evidence from a Speed Dating Experiment” research paper[2] studies the differences in various factors that influences people of various racial backgrounds to choose a partner. Their findings state that females exhibit stronger racial preferences than males. Furthermore, they observe stronger same race preferences for blacks and Asians than for Hispanics and whites. Accounting for self-reported shared interests considerably reduces the observed effect of racial preferences.

References

[1] Raymond Fisman J., Sheena Sethi Iyengar, Emir Kamenica, Itamar Simonson *Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment.* Quarterly Journal of Economics, 2006.

[2] Raymond Fisman J., Sheena Sethi Iyengar, Emir Kamenica, Itamar Simonson *Racial Differences in Mate Selection: Evidence From a Speed Dating Experiment.* The Review of Economic Studies Limited, 2008.

[3] OpenML: SpeedDating
`https://www.openml.org/d/40536`

[4] WikiPedia
`https://en.wikipedia.org/wiki/`