Data Analytics Project: Speed Dating

Caroline Nelson, Billy Yuan, Liuxuan (Kelly) Yu, Lindsay Tober

Description of Project Goals

Description of the Dataset

The Speed Dating data set was taken from a study conducted by Columbia University from October 2002 to April 2004, looking at which characteristics of 277 men and 274 women were most attractive to the opposite sex. Each subject was given a survey before participating in the experiment in which they gave demographic information, hobbies and interests, why they were participating in speed dating, and the importance of six attributes in a partner. The main six attributes that were studied were attractiveness, sincerity, intelligence, sharing interests, fun, and ambition. The data set gives, for each subject, every person they met, and their rating of that person (1-10) for each attribute. From those ratings, and how well their four-minute conversation went with each person, they decided if they'd be willing to go out with that person in the future (decision or 'like'). A pair is considered a match if they make the same decision about each other.

Some examples of characteristics that we analyzed were race, gender, career, age, and hobbies. In our analysis, we looked at which of these factors mattered in decisions, and therefore, matches. What attributes are the most desirable? What attributes or characteristics are least desirable? Do females and males desire different characteristics of partners? Many of these questions were examined in our exploratory analysis. The key questions that we completed more detailed analysis on were specifically: 'How can you increase the chances of someone 'liking' you?' and 'How can you increase the chances of someone matching with you (mutual 'likes')?'.

Importance of the Problem

We picked this data set to gain insight into what characteristics are attractive to the opposite sex. This insight could help people improve their dating chances, as long as the things needing improvement are controllable.

Exploratory Analysis

Preliminary Analysis on Data Skews

The data set was biased in a few ways. For example, the careers of the subjects were representative of those prevalent in New York, where the study took place. For this reason, we could say that a different sample could have given different prevalent career types, different personalities, and thus, different results.   It was also biased in the number of each race that participated. Over 80% of the sample were of Caucasian or Asian race. This could cause skew in personalities, and the sample as a whole. In questioning whether race mattered, subjects of these two groups, on average, claimed race mattered to them; so, a partner of a different race than these may have had a hard time finding a match.

At the same time, not everyone cared about finding a match; our analysis showed that the majority of people participating were there for the purpose of having fun and meeting new people. Whether people cared about matches or not, we have to assume that they filled out their survey honestly. The surveys showed that people were more likely to have between five and ten 'likes' than more than ten; this was another skew in the data. With the characteristics mentioned in mind, we went on to investigate reasons for desirability and compatibility.

Data Exploration

The first potential confounding variable we looked at was the order percentile at which a pair met, whether they were paired first or last in their wave. It showed that the orders were just randomly moving according to the order percentile, so order was not a confounding variable.

Looking at activities, we saw that sports-related activities and exercise had a high correlation, so we grouped them together as outside hobbies; another group was art hobbies consisting of museums, art, theater, movies, concerts, music and reading. We expected that people with similar hobby scores would have higher match rates. However Figures 1 & 2 show that there is no significant difference in match rates. This may be because the 'date' only lasts four minutes, so they could have had the same interests but didn't draw that conclusion from a four-minute conversation.

When exploring race, pairs of the same race have a slightly higher match rate than pairs of different races. Regardless of race however, people had close to a 40% chance of getting a 'like'. Being open to dating other races may also vary across race, but the skewness in the data set did not allow for a detailed comparison.

Disregarding gender, when we looked only at age differences, we found that people tended to match more frequently with others closer to their age (Figure 3). When we separated female and male to visualize the relationship between age differences and match rate, female decisions were not influenced much by age. However, when females' ages were higher, males tended to reject them more.

Another methodology for exploring the importance of attributes is to look at the percentage of matches by attribute scoring pairs. This takes the percentage of matches against total partners for each combination of attribute ranking pairs (Figure 4). These heat maps were built using the cross-tabulation for each partner's ratings for each other, by pairing, then calculating the percentage that were a match. The heat maps give a visual depiction of trends and correlations across attributes, showing disparities between men and women's ratings.

Regarding attractiveness, we can see that men were more selective than women because women with higher attractiveness ratings matched with men who had lower attractiveness ratings much more frequently than men who had higher attractiveness ratings did with women who had lower attractiveness ratings. This validates the 'double standard' that is often complained about in culture, potentially seen most explicitly in the woman who was a '10' matched with a '2' man whereas no man who was a '10' matched with a woman below a '4'.

<u>Solution and Insights</u>

Logistic Regression / Naive Bayes to Predict Getting a 'Like'

Demographic factors such as race and career did not seem to affect whether someone received a 'like.' On the other hand, attributes such as attractiveness and shared hobbies seemed to have a positive effect. In order to validate these hypotheses, these variables were applied to both logistic regression and Naive Bayes models; variables with higher coefficients would be more important factors.

After performing multiple iterations for both the logistic regression and Naive Bayes, the most important variables for both men and women seem to be consistent: attractiveness, fun, and shared interests (Figure 5). Using only the attributes and race in the models yielded a test set accuracy of 75%, compared to a 62% base accuracy for the negative class. Based on coefficient weight, the logistic regression shows that a woman's attractiveness is much more important than other factors for men, whereas the coefficients of the top 3 attributes were more evenly spread for women.

However, the logistic regression and Naive Bayes models seemed to differ in the magnitudes of these coefficients. The coefficients of the most important variables are evenly distributed for both men and women; this contrasts with what the logistic regression showed for men. The even distribution may be due to how Naive Bayes considers each variable independent, making the model susceptible to collinearity among variables.

Logistic Regression to Predict Getting a 'Match'

Logistic regression was also performed on 'matches' based on the six rated attributes (Figure 6), which validated some of the conclusions found in the logistic regression on 'likes.' Based on coefficients, the model indicated that men still emphasized attractiveness, but it was much more evenly distributed with fun and shared interests given these were matches instead of just likes. For women, fun was the most important indicator from a coefficient standpoint, followed by shared interests and attractiveness.

Running an additional logistic regression on matches including other variables such as attitudes, age, and race revealed some additional information. Women preferred men who were looking for a serious relationship and who went on dates around once a week. One of the biggest turnoffs for women was men speed dating just to say that they did it. On the other hand, men preferred women speed dating just to say that they did it, along with women they found attractive. One of the biggest turnoffs for men was women of a different race, which could be skewed by the data set.

K-Means Clustering to Determine What 'Matches' Have in Common

What are characteristics of 'matches' and how could they help explain why two people "like" each other? The previous logistic regression and Naive Bayes models explained factors that were important for getting someone to like you. K-Means may help explain why two people are a match. The data was restructured so that the interests and attributes of both men and women who matched together were in the same row.

Four of the clusters turned out to be somewhat relevant (Figure 7). With the exception of sports, shared interests do not seem to matter too much. One cluster contained matches where the men had eclectic interests but the women's hobbies were not relevant. This is no surprise given that each man-woman pair only had four minutes to get to know one another. Perhaps sports is an easy topic to bond over, and these matches did not get a chance to explore other interests.

Conclusion

While the analysis revealed a number of variables that help increase chances of desirability and compatibility (e.g., attractiveness, fun), skews in the data and limited match information prevent extrapolating our conclusions to the population at large without further validation.

Appendix

Figure 1

Hobby Influence on Matches: “Outside” Hobbies

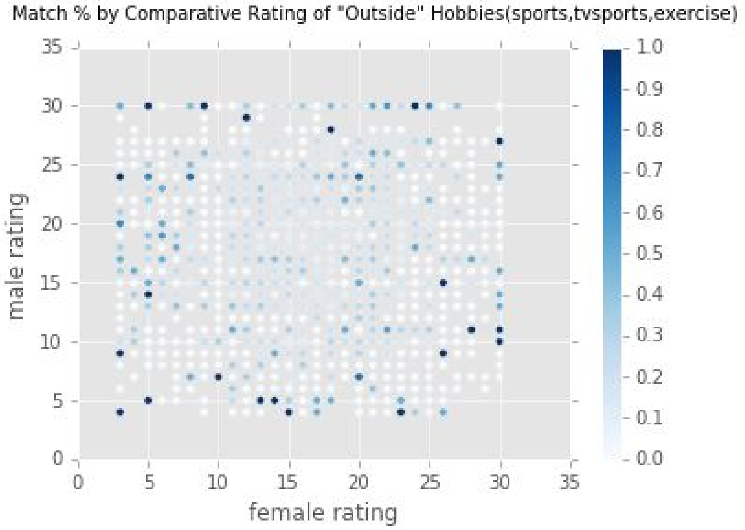


Figure 2

Hobby Influence on Matches: “Art” Hobbies

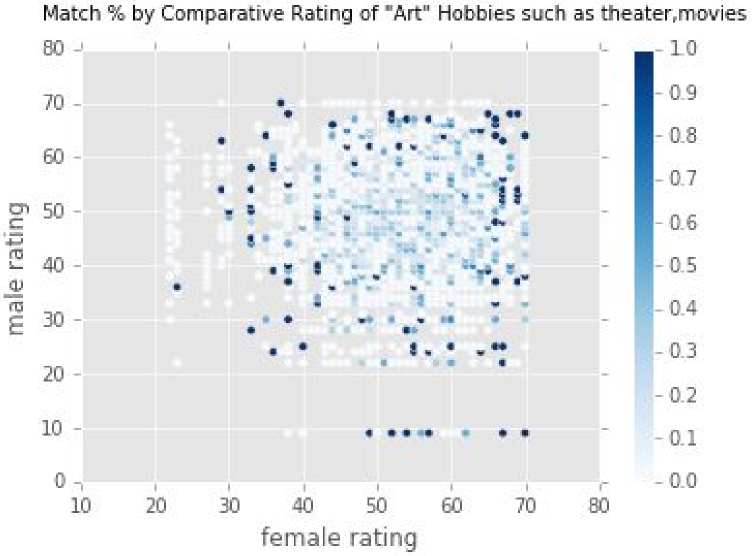


Figure 3

Relationship between Age Difference and the Average Match Rate

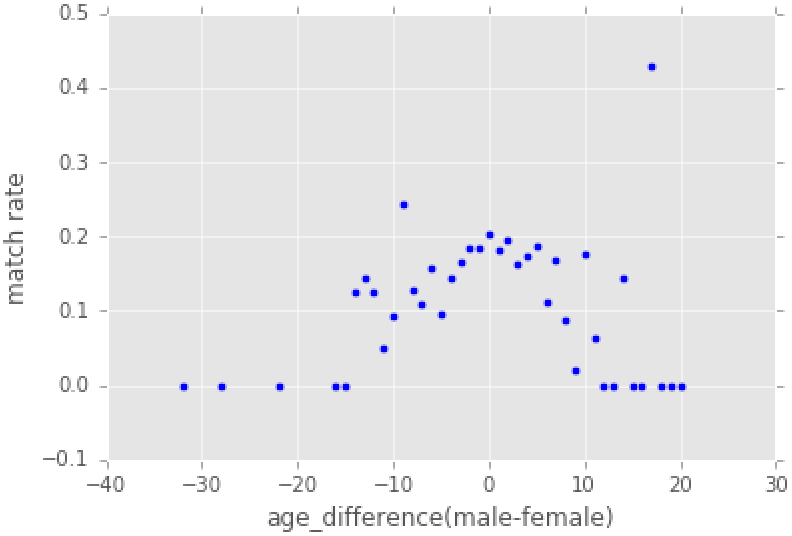
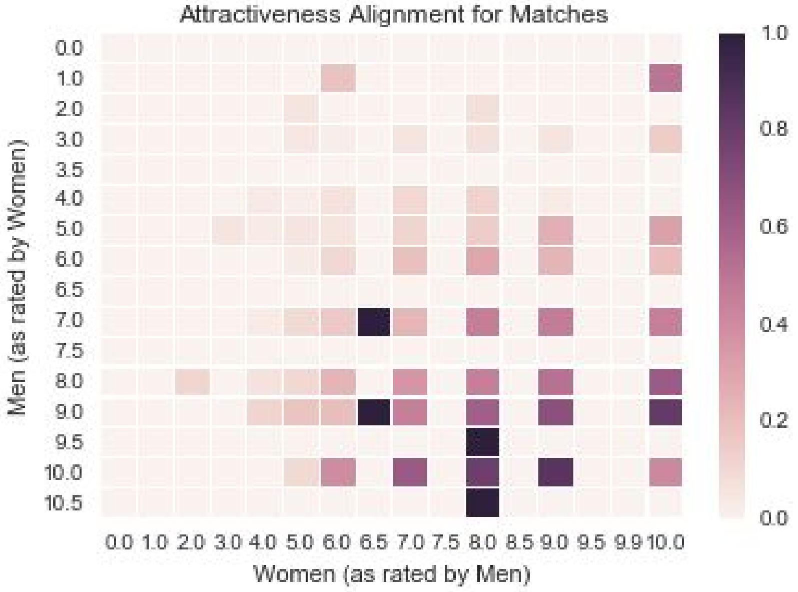
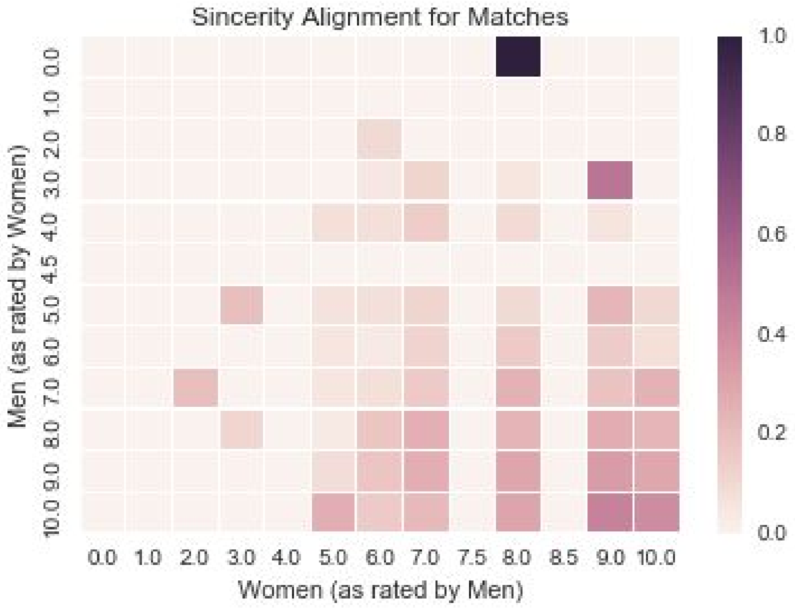
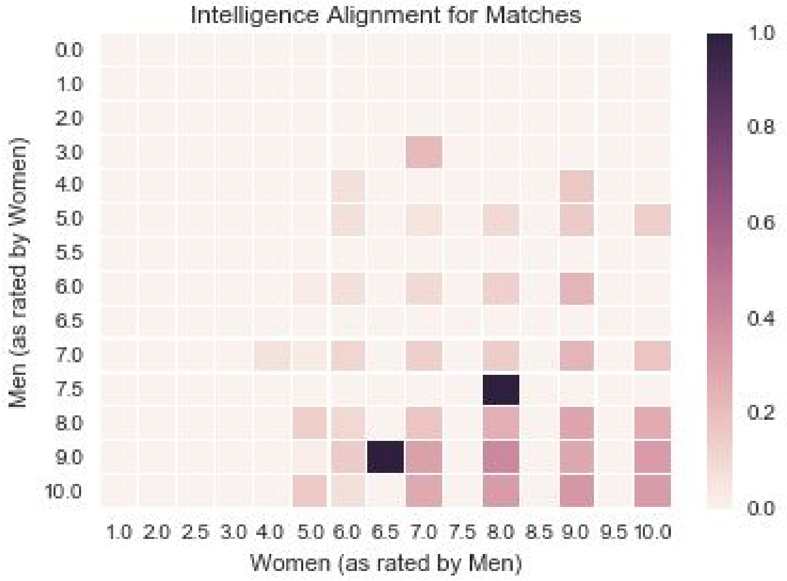


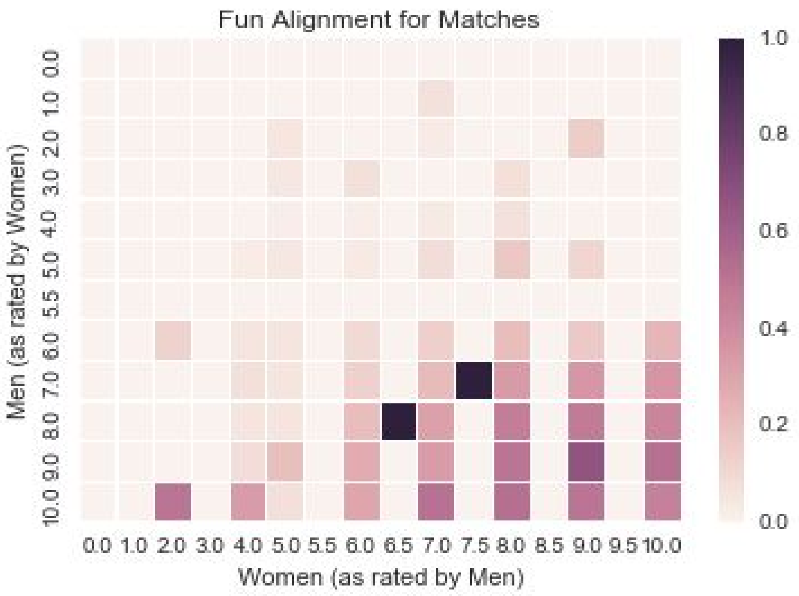
Figure 4

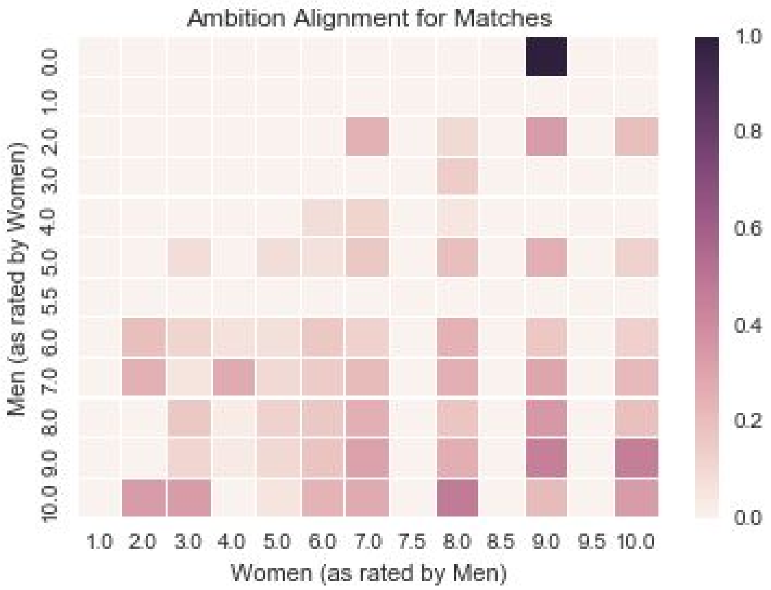
Attribute Heatmaps




Sincerity Alignment for Matches


Intelligence Alignment for Matches


Fun Alignment for Matches
Men (as rated by Women)
0.0
1.0
2.0
3.0
4.0
5.0
5.5
6.0
7.0
8.0
9.0
10.0
0.0 1.0 2.0 3.0 4.0 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0 9.5 10.0
Women (as rated by Men)
1.0
0.8
0.6
0.4
0.2
0.0


Ambition Alignment for Matches
Men (as rated by Women)
0.0
1.0
2.0
3.0
4.0
5.0
5.5
6.0
7.0
8.0
9.0
10.0
1.0 2.0 3.0 4.0 5.0 6.0 7.0 7.5 8.0 8.5 9.0 9.5 10.0
Women (as rated by Men)
1.0
0.8
0.6
0.4
0.2
0.0

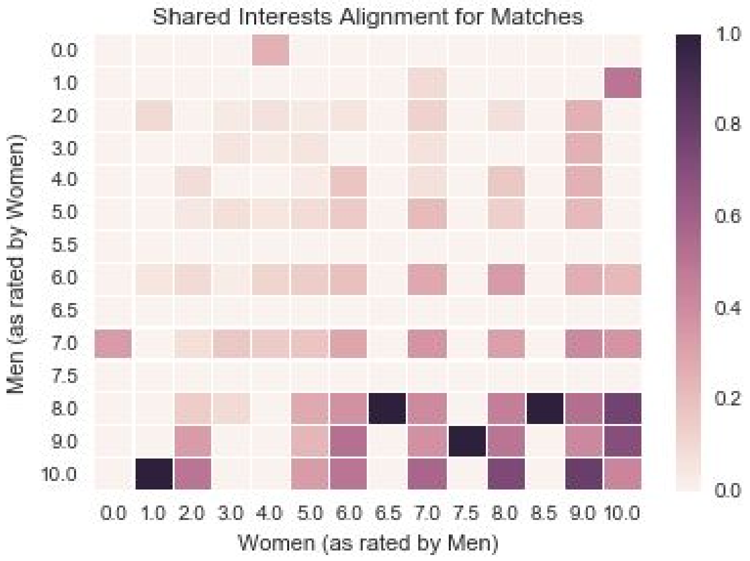


Figure 5

Logistic Regression Coefficients

Variable	β(W->M)	β(M->W)
Attractive	.374	.663
Fun	.290	.264
Shared Interests	.243	.277
Race (white)	.351	-.110
Intelligence	.082	-.037

Naive Bayes Weights

Condition	**W->M**	**M->W**
Shared interests >= 7	1.000	.970
Attractive >= 7	.962	1.019
Fun >= 8	.753	.740
Race (white)	.286	.176
Intelligence >= 7	.258	.259

Figure 6

Logistic Regression *(Attributes)*

Variable	β(W->M)	β(M->W)
Attractive	.164	.248
Fun	.272	.245
Shared Interests	.235	.227
Intelligent	-.041	.034
Sincere	-.003	-.119
Ambitious	-.145	-.137

Baseline accuracy (men): 80.9% (negative class)
Baseline accuracy (female): 81.3% (negative class)

Figure 7

K-Means Clusters

Attributes are in the following format: attribute - gender. For example, “Confidence - men” means that men who scored high in ‘confidence’ were grouped together.

Top 4 Variables per Cluster

Cluster 1 - Likable and confident	**Cluster 2 - Eclectic Interests**	**Cluster 3 - “Mysterious Artists”**	**Cluster 4 - Sport enthusiasts**
Confidence - men	Likes Concerts - women	Likes art - men	Likes sports - men
Likability - women	Likes music - women	Likes museums - men	Likes sports - women
Fun - women	Likes museums - women	Likes theater - men	Likability - men
Shared interest - women	Likes art - women	Likes concerts - men	Likes watching sports - men