QuickMatch Analytics: Decoding Success in Speed Dating

Group 14: Adam Kaiser, Lindsey Myers, Viraj Shah, Christine Yoo

December 12, 2023

Introduction

The dataset utilized in our project originates from a Speed Dating event involving students from Columbia University's graduate and professional schools. Participants were enlisted through mass emails, distributed flyers, and promotional efforts by research assistants. During the speed dating session, each participant engaged in four-minute interactions with counterparts of the opposite sex. The order and session assignments were randomly allocated. Following each speed date, participants completed a form, assessing their date on a scale of 1-10 across various attributes. It's important to note that the dataset exclusively includes data from the initial date within each session.

To further inform, the participants gave a ranking 1 to 10 based on certain traits along with stating if they would go on another date. The number 0 meant no to a second date and 1 meant yes. Some of the traits included attractiveness, sincerity, intelligence, amount of fun, ambitiousness, and shared interests. The dataset also included a ranking on how much the person liked the other person as well as the individual's age and race. Another question the participants had was how much they thought their date liked them. They gave this a rating of 1-10 as well.

Research Aims and Methods

For this project we used the information given in the dataset to find if males and females differ in how much attractiveness influences their decision for another date and what trait was most significant. To answer these questions, we used a variety of different methods. This includes finding the average rating of each trait, looking at the distributions of each trait, linear regression models, and knn. We also asked the question: "Can predictive models be used to determine if an individual has a chance of a second date occurring based on these ratings?"

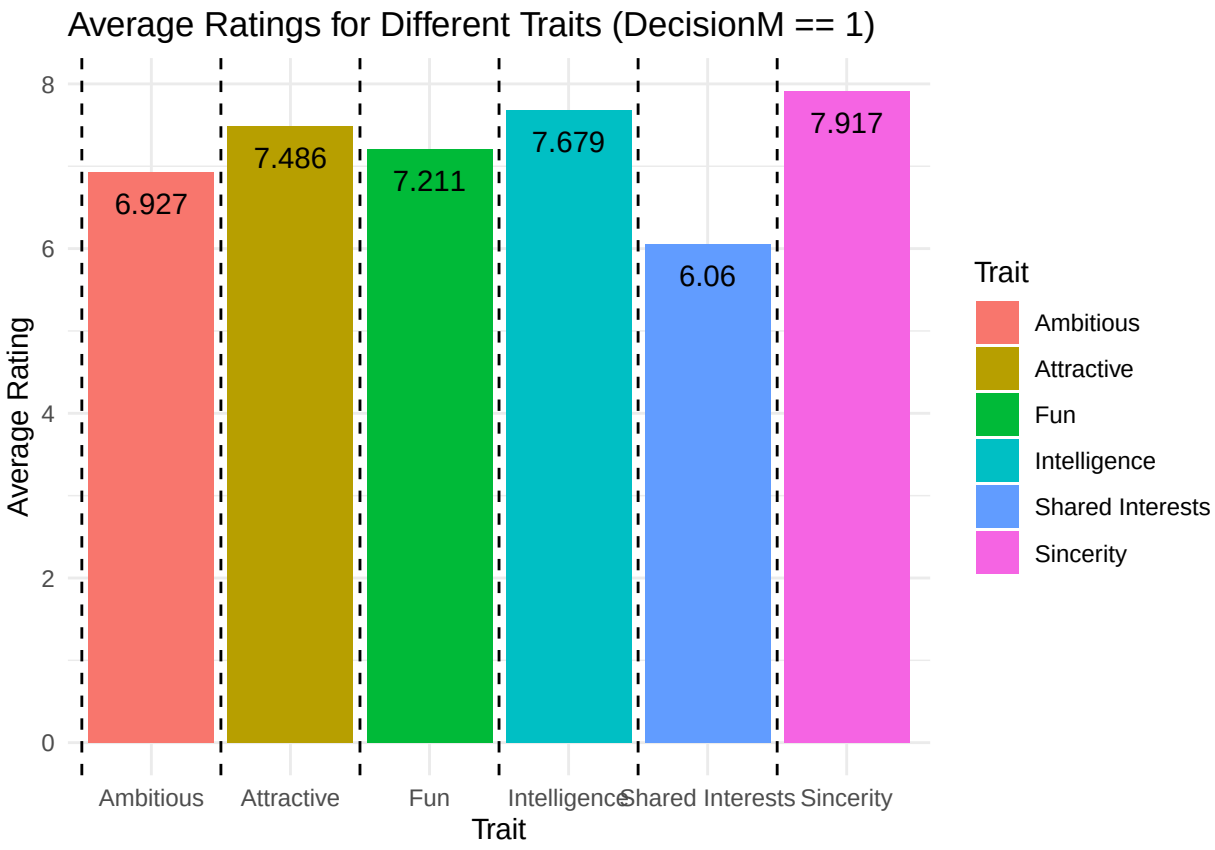
For the first question: "Do males and females differ in how much attractiveness influences their decision for another date?" Our group anticipated that attractiveness would be more important to males than females in the dating world based on what we experience in our personal lives. To answer the second question: "What trait is most significant?" We thought that attractiveness would be the most significant trait for males and fun would be most significant for females. We also anticipated that predictive models can be used to determine if an individual has a chance of a second date occurring based on the ratings in the dataset.

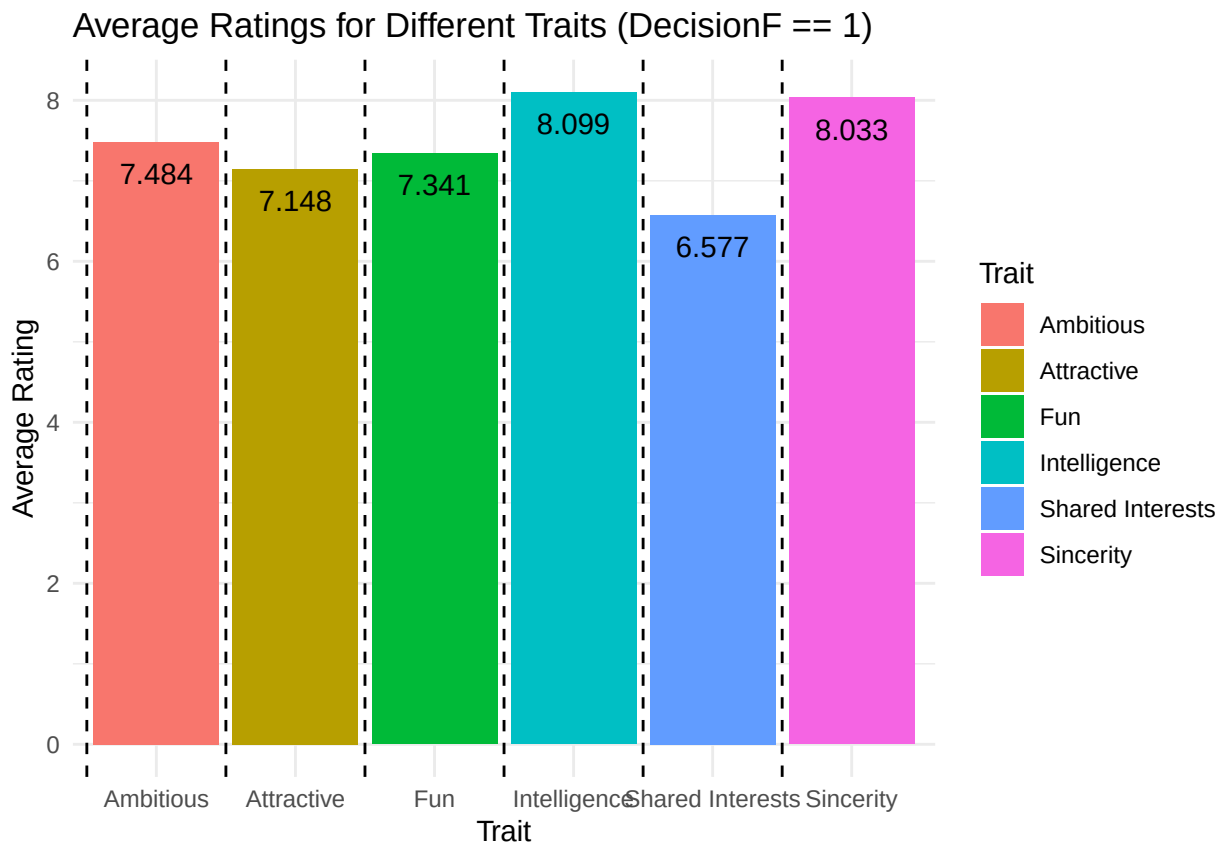
Results

For our first question, our group mainly used techniques related to subsetting and filtering the original dataset. We first looked at the cumulative averages of each trait, given that a second date occurred. We did this for both sexes by filtering for a decision value of 1 for each, relating to a positive experience overall.

Upon examining each graph, we see the following.

1. Which trait is most important when predicting if a second date will occur? Does gender affect which trait is most important?



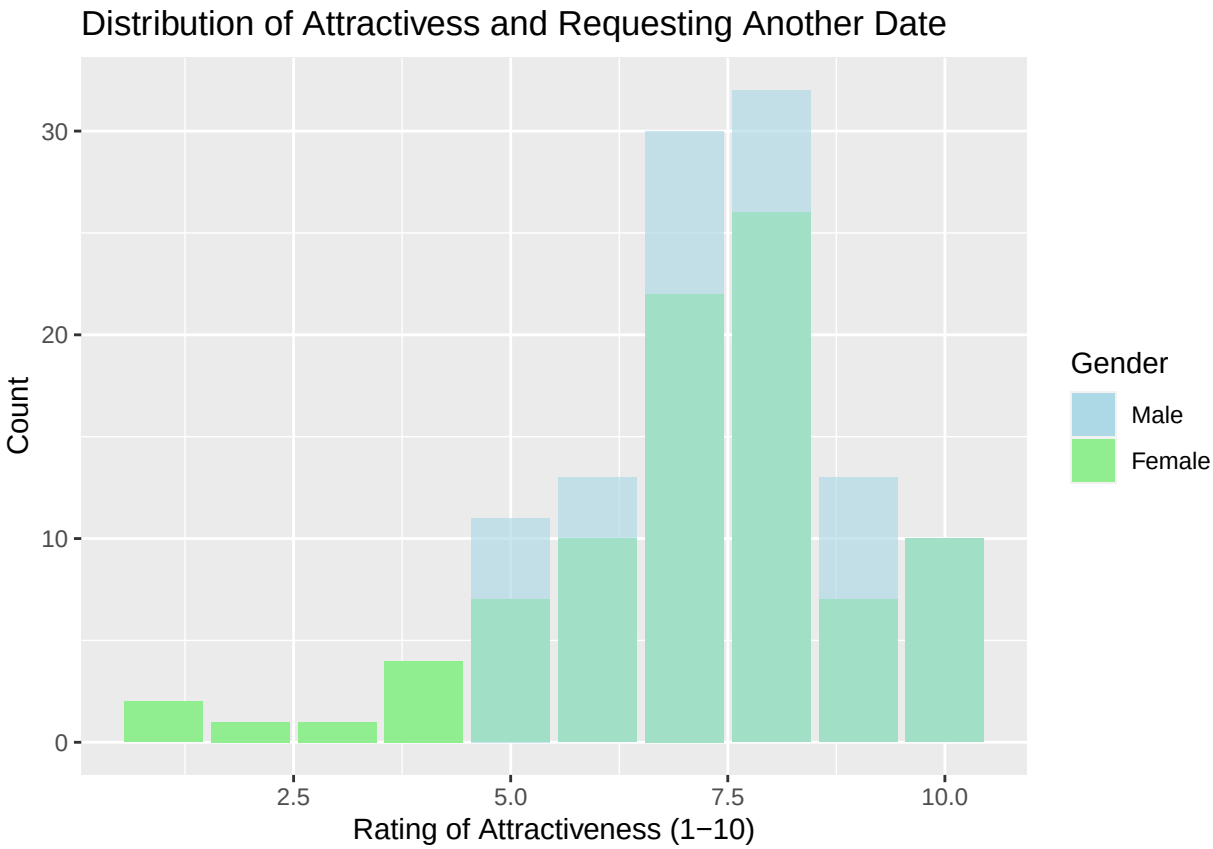


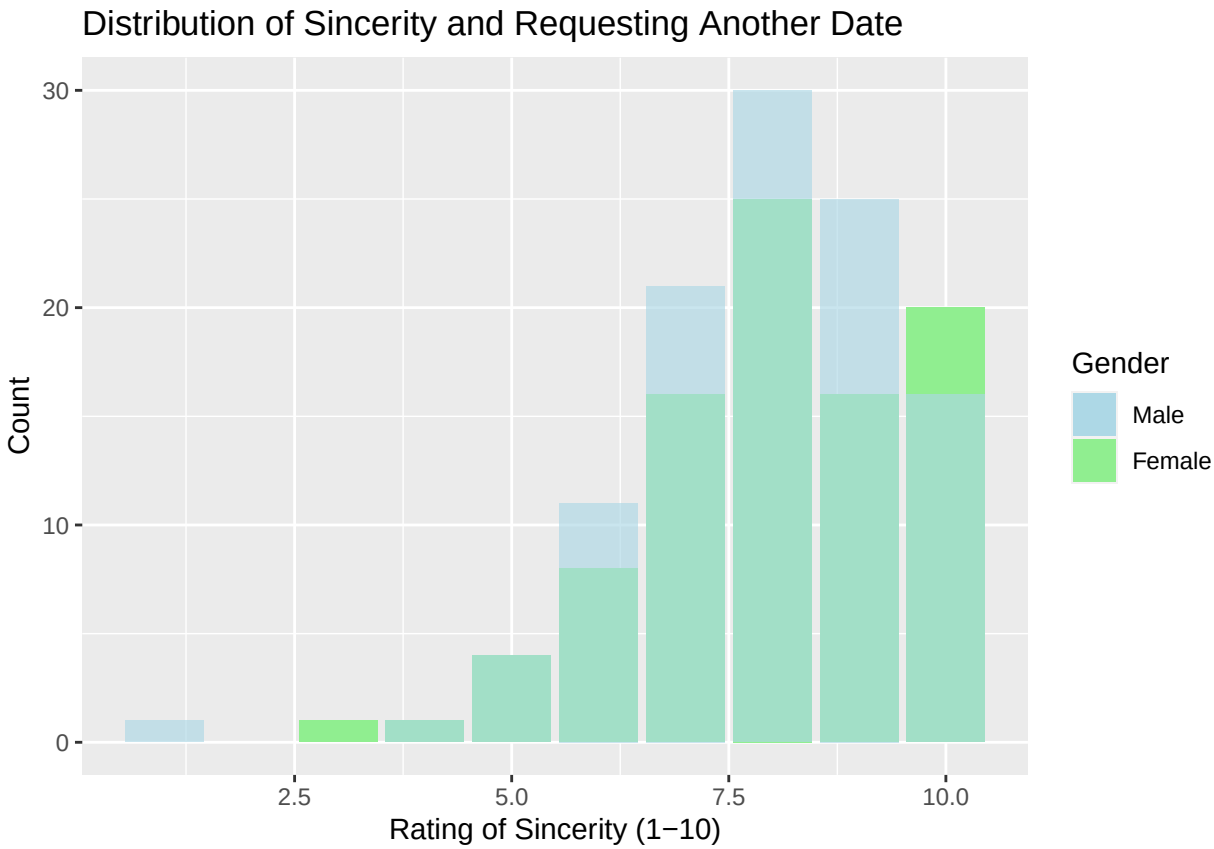
The average ratings for the male model show that sincerity has the highest average rating, which suggests that among the presented traits, it is considered the most favorable or important. Shared interests has the lowest average rating, indicating it is the least favored or important trait in this context. The other traits (Attractive, Fun, Intelligence, Ambitious) have relatively similar average ratings, falling between the highest and lowest values. The average ratings for all traits fall above the mid-point of the scale, indicating a generally positive rating for all traits. Overall, the plot shows a ranking of traits in terms of their average perceived importance or favorability by the respondents (males who chose to go on a second date), with Sincerity being the most valued and Shared interests being the least valued.

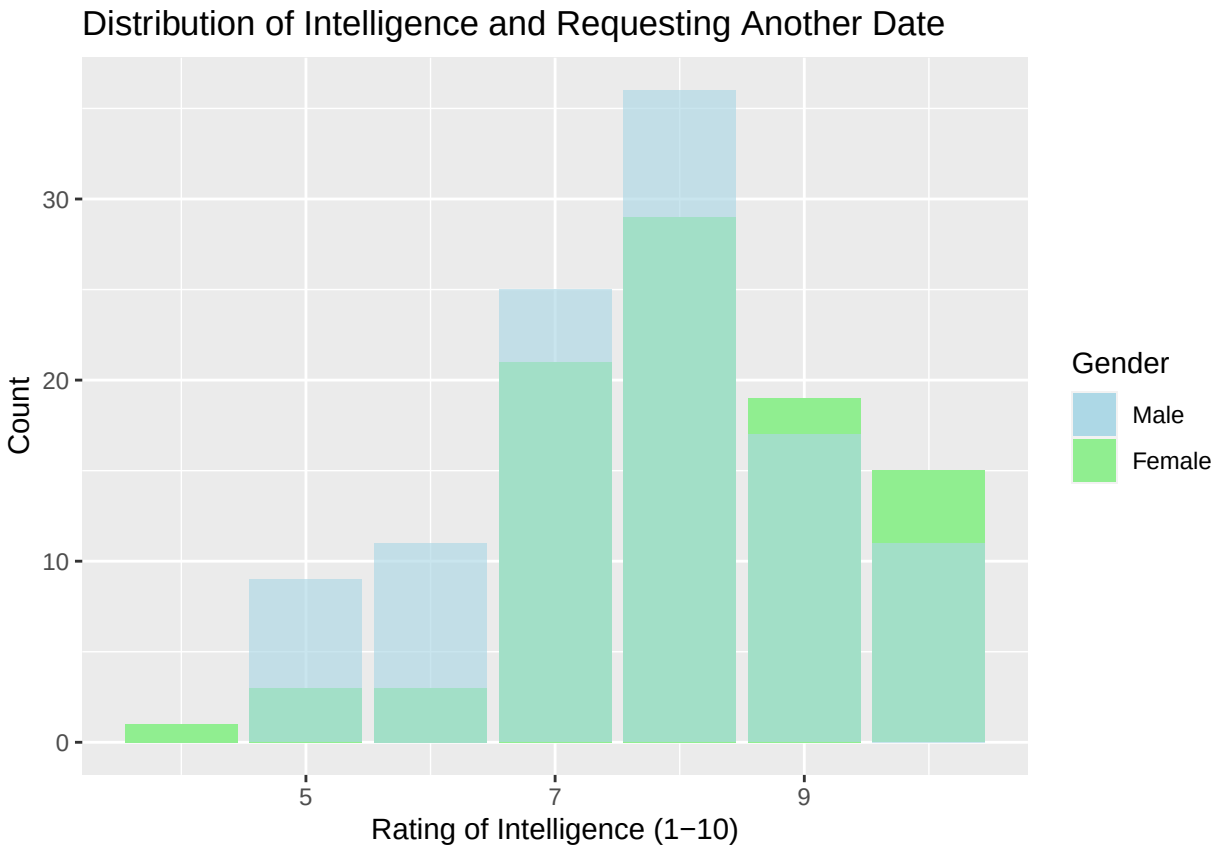
The average ratings for the female model show that intelligence has the highest average rating, suggesting it is considered the most favorable or important trait for the female-decision context as well. Shared interests was the trait with the lowest average rating, indicating that it is perceived as the least important or favorable trait among those listed. The ratings for Attractive, Fun, Sincerity, and Ambitious are very similar to each other, and like in the previous plot, they are positioned in the middle of the scale, implying a moderate level of importance relative to the others. The distribution of ratings in this plot is almost identical to the distribution in the prior plot, which suggests that the perceived importance of these traits does not significantly differ between the "DecisionM" and "DecisionF" conditions, at least based on the average ratings. All traits have an average rating above the mid-point of the scale, which indicates a positive evaluation overall for each trait.

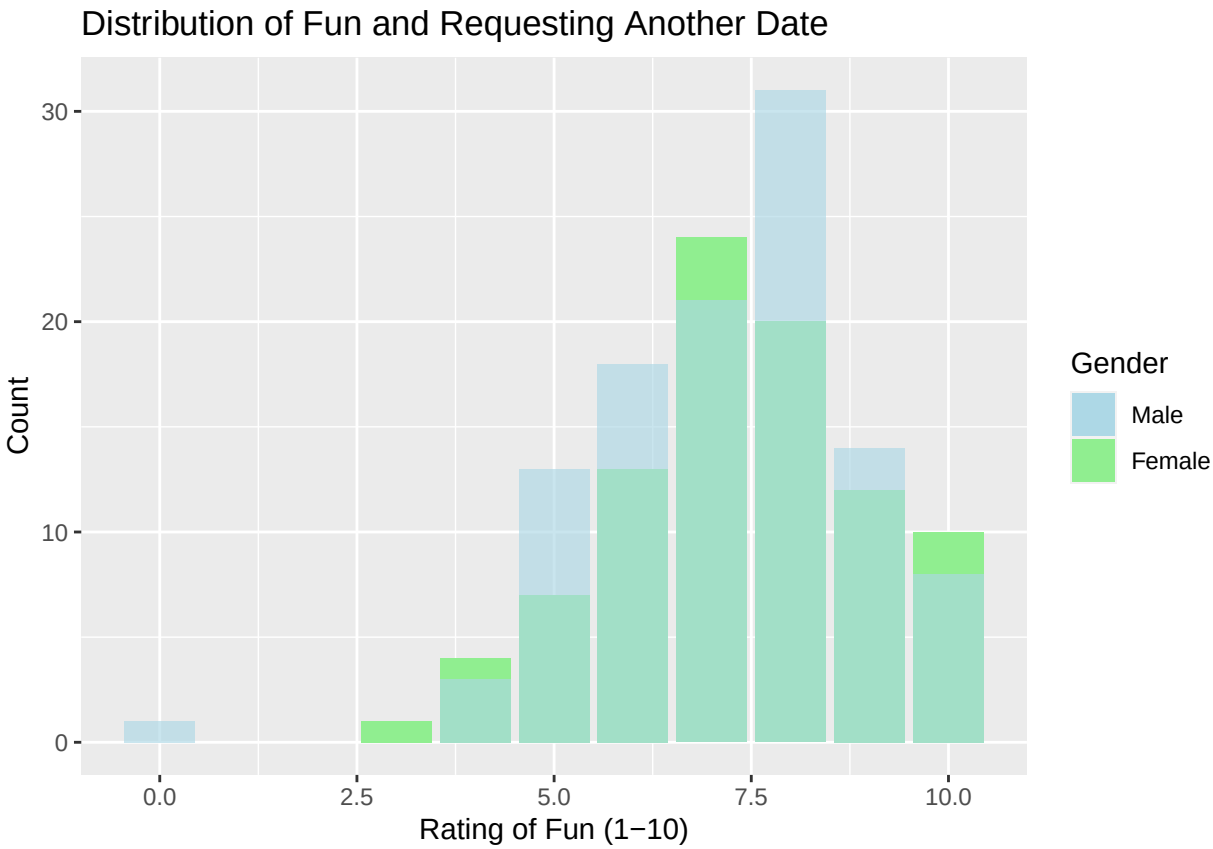
This provided a solid preliminary foundation for what we might expect to influence each decision level (date or no date, 1 or 0), which was further explored by examining the specific distributions of each trait.

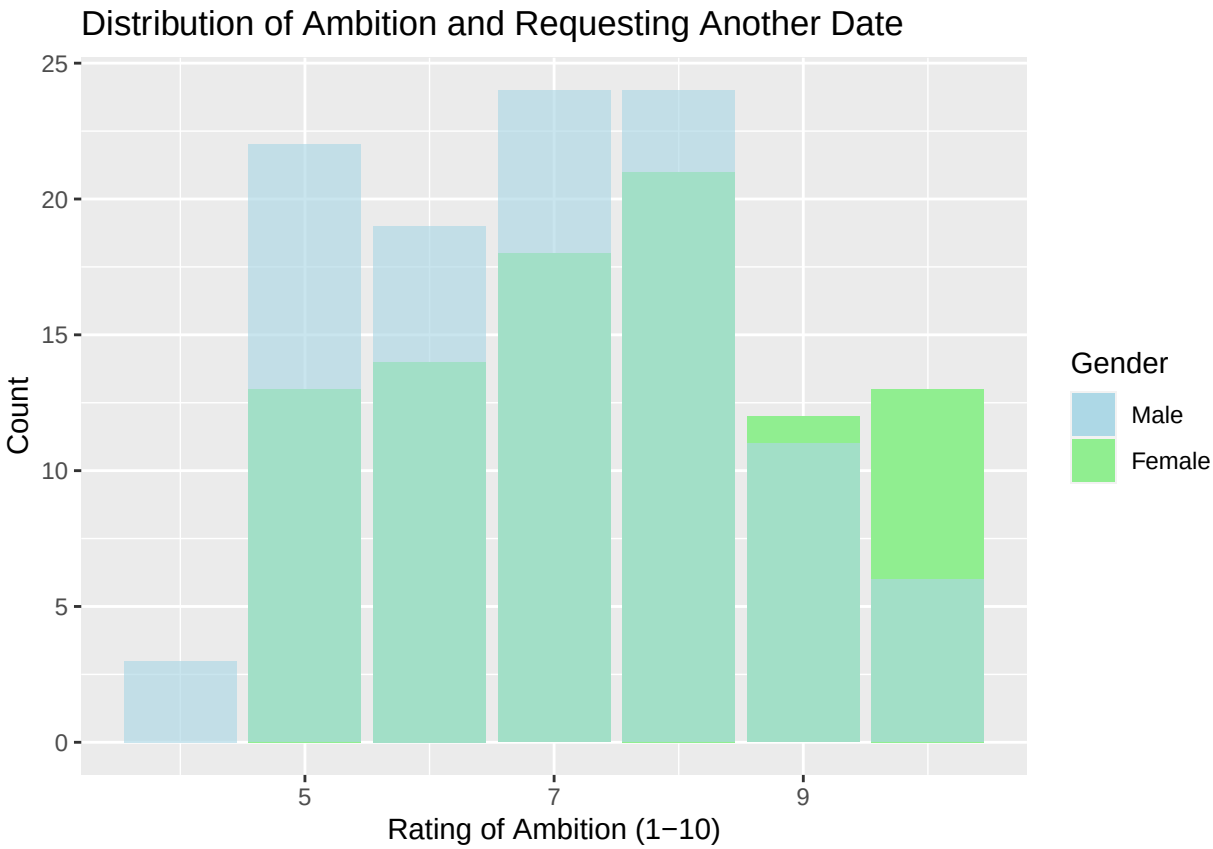
As stated, we then examined the distribution of each trait to see if there were any trends which might enlighten us to other exploratory steps. So, we again subsetted each trait such that we only had values for a confirmed second date (decisionM or decisionF = 1). Then, we plotted the counts of individuals' rating levels for each trait.

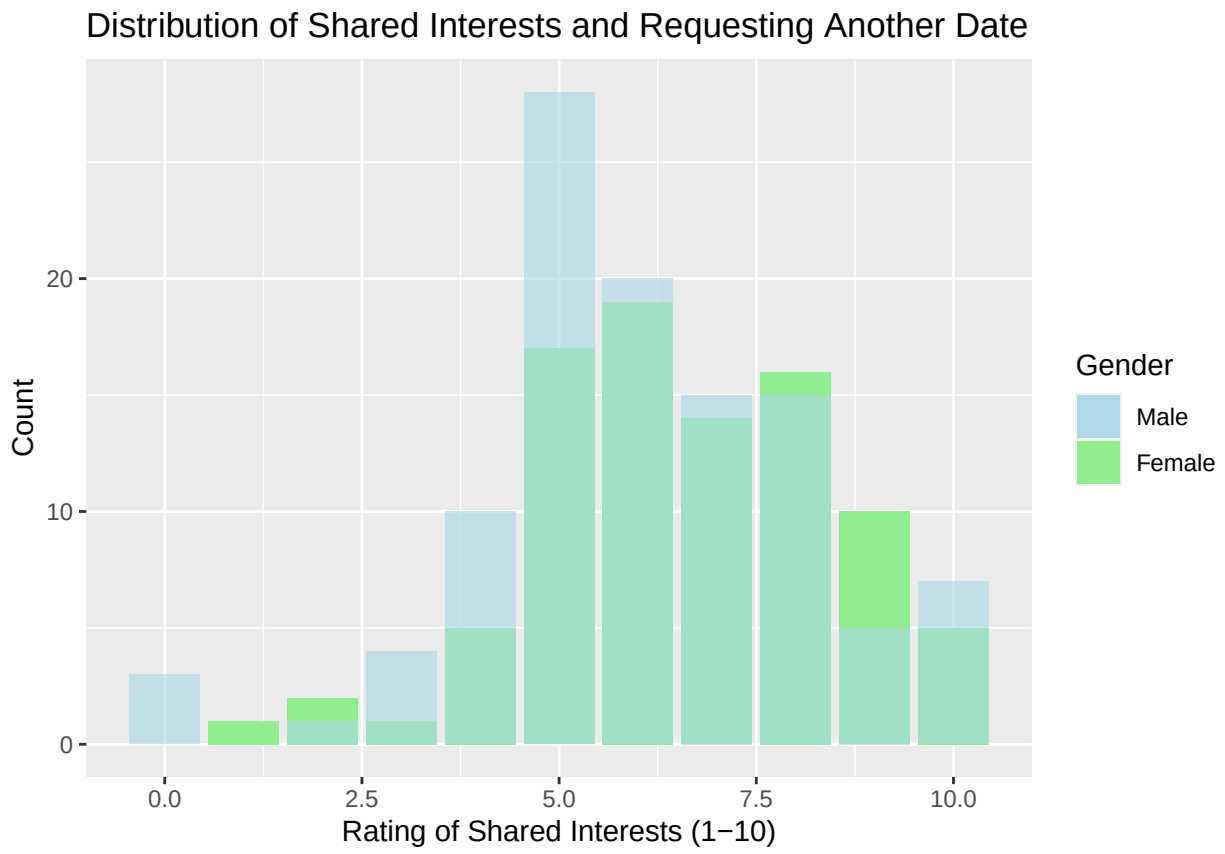

Distribution of Attractivess and Requesting Another Date
Count
30
20
10
0
2.5
5.0
7.5
10.0
Rating of Attractiveness (1−10)
Gender
Male
Female


Distribution of Sincerity and Requesting Another Date
Count
30
20
10
0
2.5
5.0
7.5
10.0
Rating of Sincerity (1−10)
Gender
Male
Female


Distribution of Intelligence and Requesting Another Date
Count
30
20
10
0
5
7
9
Rating of Intelligence (1−10)
Gender
Male
Female


Distribution of Fun and Requesting Another Date
Count
30
20
10
0
0.0
2.5
5.0
7.5
10.0
Rating of Fun (1−10)
Gender
Male
Female

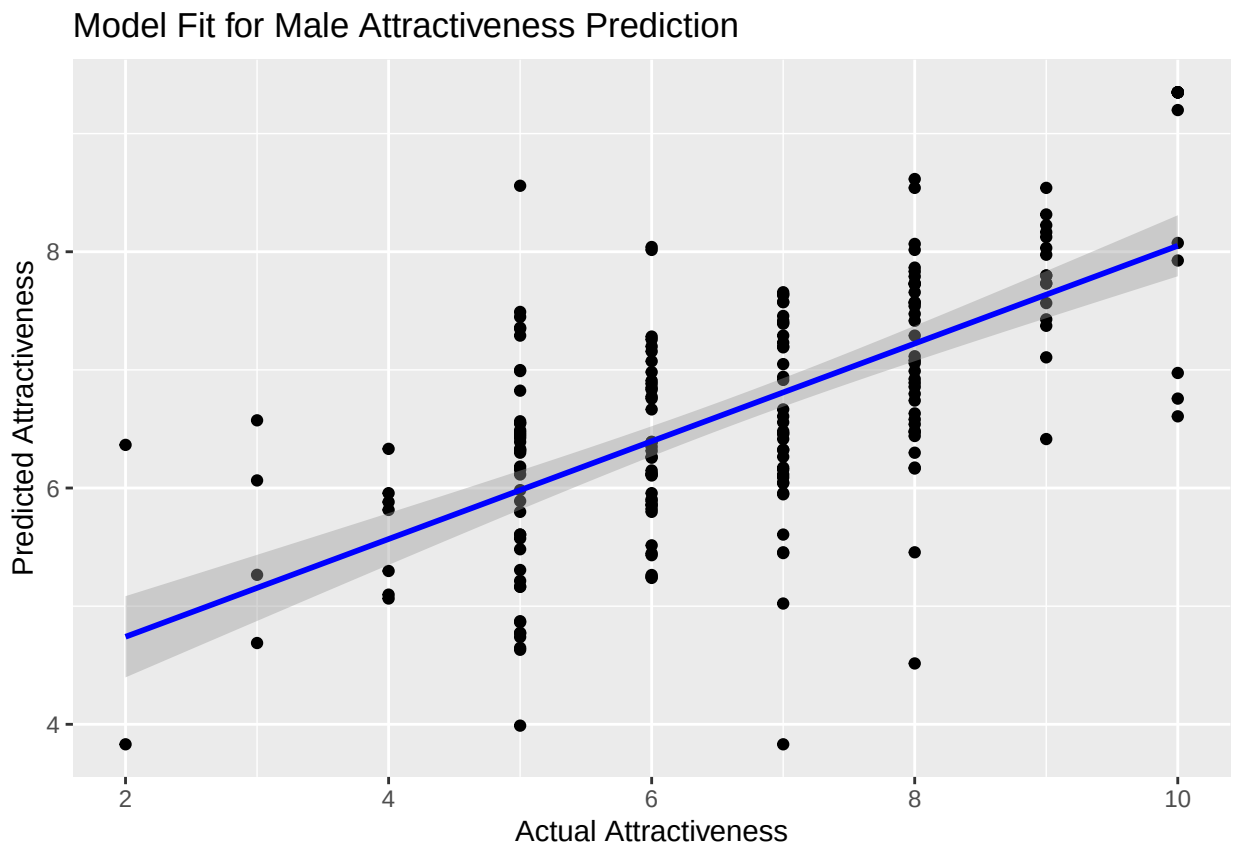

Distribution of Ambition and Requesting Another Date
Count
25
20
15
10
5
0
5
7
9
Rating of Ambition (1−10)
Gender
Male
Female



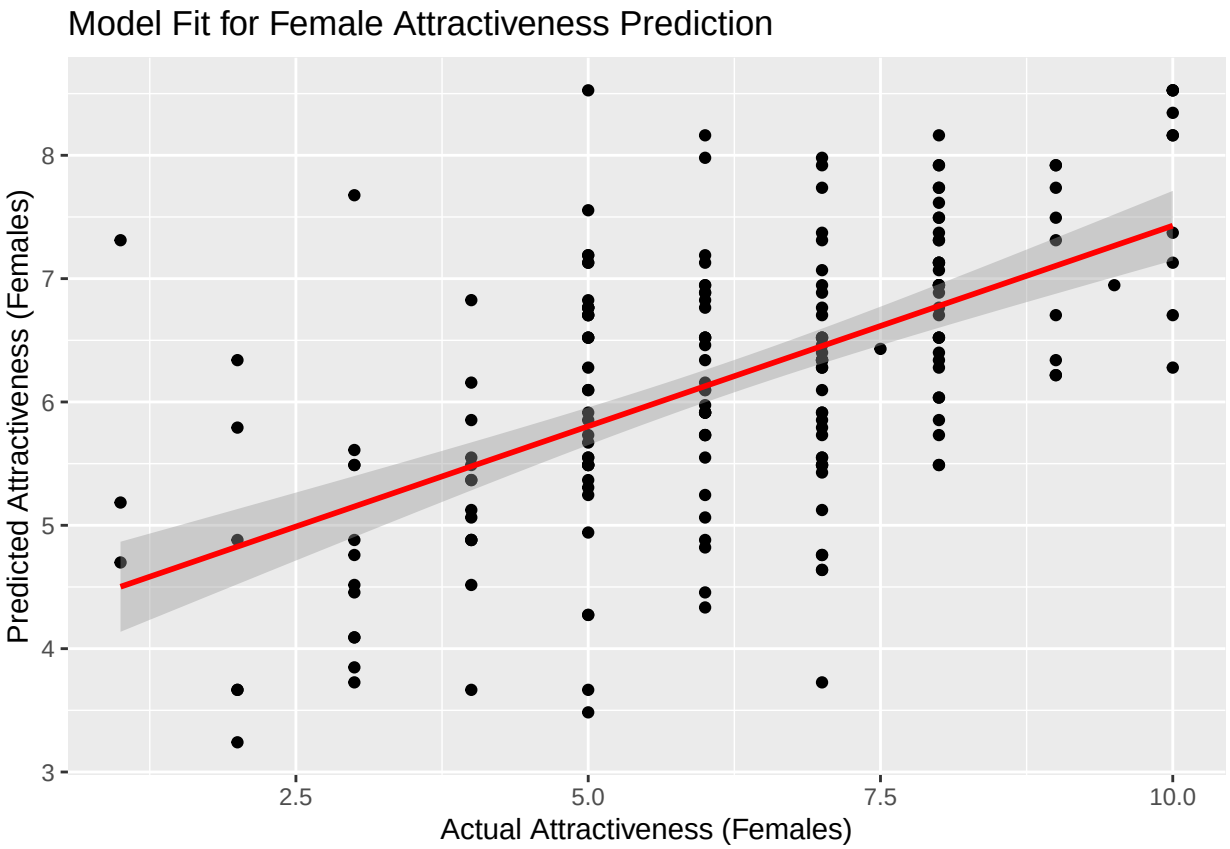
Upon examining the graphs for each trait we see that all plots are left skewed except for the plot for ambition. We can see that for attractiveness, males did not want a second date with females that were rated under a 5. There was a low number of females wanted to go on a second date if the male was rated under a 5. This suggests that males let attractiveness influence their decision on a second date more than females. The plot for ambition had the most random distribution of all the traits. Overall, females rated ambition higher than males did. The plots for intelligence, fun, and sincerity were all relatively the same between males and females.

2. Can predictive models be used to determine if an individual has a chance of a second date occurring based on these ratings?

After examining the distributions of averages, as well as the distributions for each respective trait, we then decided to create a linear regression model for predicting attractiveness from the other five traits. Upon doing this, we displayed the results inside of linear regression plots, with one model for each sex.



The figure shows a positive linear relationship between actual attractiveness and predicted attractiveness for males, indicating that as the actual attractiveness rating increases, the predicted attractiveness also tends to increase. The spread of data points around the regression line suggests there is some variability in the predictions, with a larger spread observed at higher values of actual attractiveness. The shaded area represents the confidence interval for the regression line, showing where we expect the true regression line to fall, with a greater degree of uncertainty at the extremes of actual attractiveness.

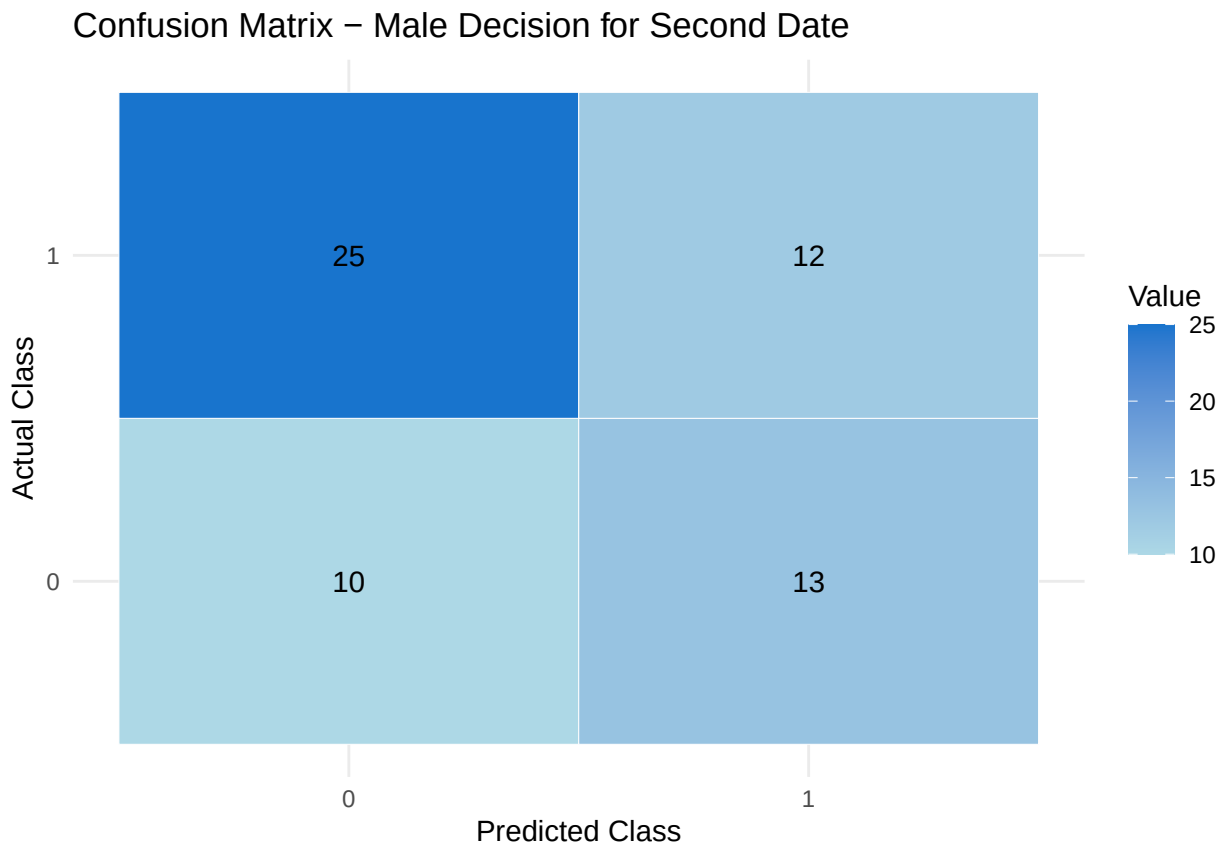


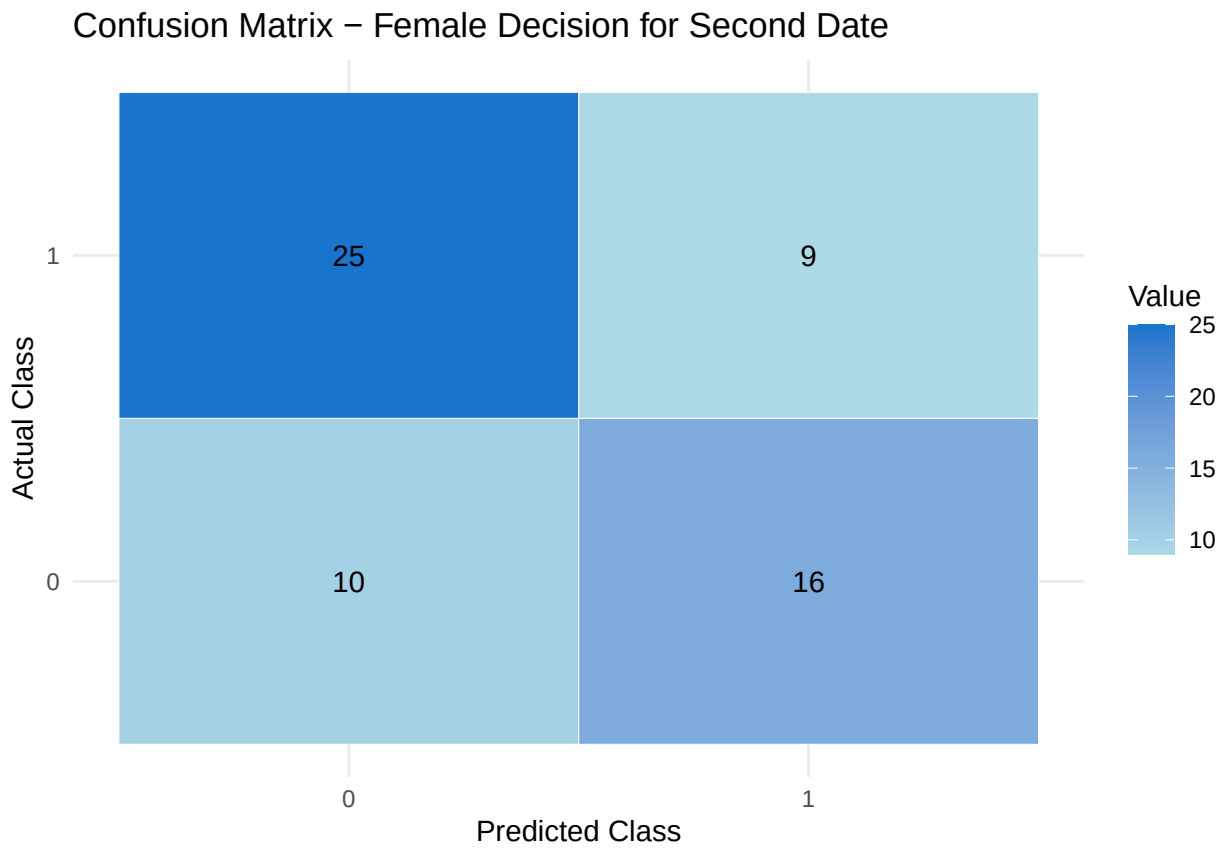
In the regression plot for female attractiveness prediction, there is a clear positive linear relationship between actual and predicted attractiveness, indicating a tendency for higher actual attractiveness ratings to correspond with higher predicted values. The data points are fairly evenly distributed around the regression line, which suggests the model's predictions are consistent across the range of actual attractiveness values. The confidence interval, indicated by the shaded area, is narrowest in the middle of the actual attractiveness range, implying more certainty in the model's predictions in this region. There are some potential outliers, particularly at the higher end of actual attractiveness, where the data points are more spread out.

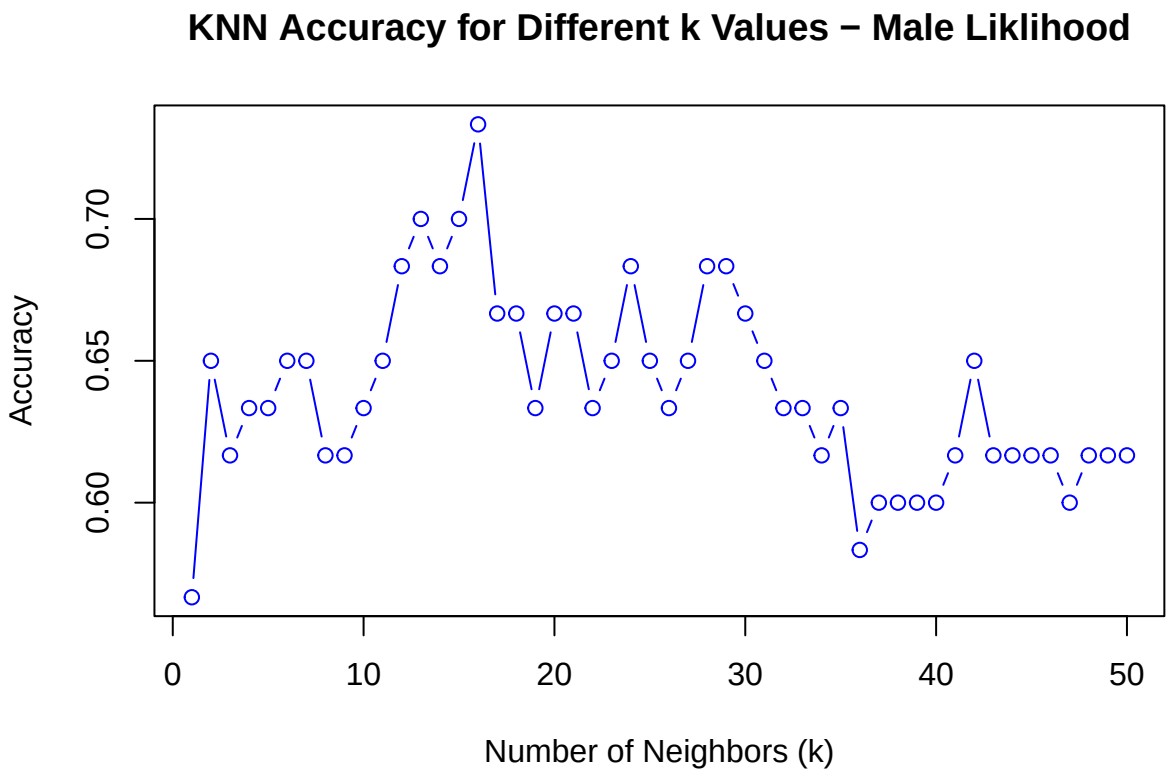
Confusion Matrix for KNN Classifier In addition to the regression model, we also decided to frame the investigation in terms of a classification problem. We decided to investigate the extent to which we could predict the likelihood of a second date, from the six independent variables/features mentioned previously. We also added the age of the participant, and the age of the partner which they were evaluating. So in summary, we wanted to determine if we could effectively predict a second date from 8 total features relating to attractiveness. To do this most effectively, we first performed iterative feature elimination for each model.

For the male model, we found that the most significant features were Like (general rating of experience), PartnerYes (estimate that partner likes them), Age (of female), Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. This allowed for the model to have an accuracy of 0.733.

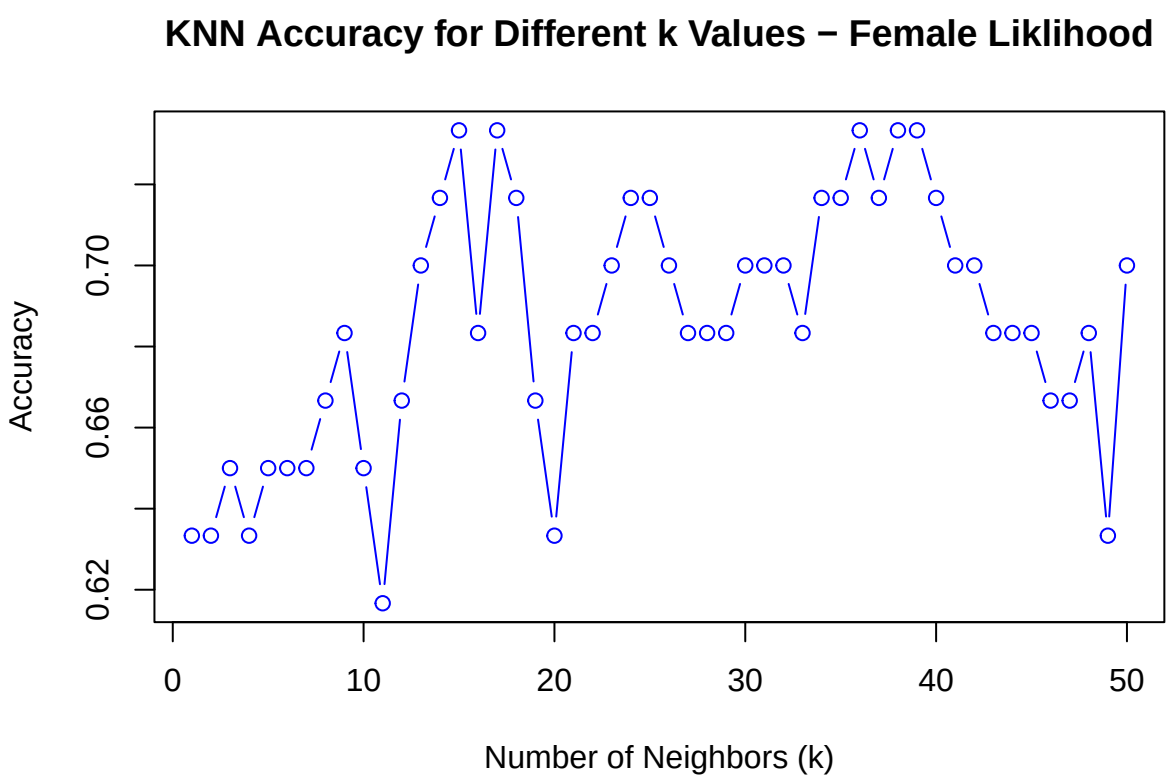
For the female model, we found that the most significant features were Like (general rating of experience), PartnerYes (estimate that partner likes them), Age (of female), Age (of male), Attractiveness, Intelligence, Fun, Ambition and Shared Interests. This allowed for the model to have an accuracy of 0.683.


Confusion Matrix − Male Decision for Second Date
Actual Class
1
0
25
12
10
13
0
1
Predicted Class
Value
25
20
15
10


Confusion Matrix − Female Decision for Second Date
Actual Class
1
0
25
9
10
16
0
1
Predicted Class
Value
25
20
15
10



KNN accuracy for different K values (Male decision): This graph depicts the accuracy of the K-Nearest Neighbors (KNN) classifier at various values of 'k', which is the number of neighbors considered for making the classification decision. The classifier(s) are predicting the male decision for a second date, and there is a fluctuating trend suggesting that accuracy varies with different 'k' values; notably, accuracy peaks at certain points (such as near k=5 and k=25), but overall, there isn't a clear upward or downward trend. The variability indicates that for this classification task, the optimal 'k' value for maximum accuracy might be context-dependent and choosing the right 'k' is crucial for the model's performance. In the context of the investigation, it was decided that a K value of ~16 was optimal for increasing performance while minimizing overfitting.



KNN accuracy for different K values (Female decision): This plot illustrates the accuracy of the K-Nearest Neighbors (KNN) classifier for varying 'k' values, but this time for the females' decision for a second date. Similar to the previous plot, the accuracy fluctuates significantly as 'k' changes, with no clear trend towards improvement or deterioration over the range of 'k' values. It shows peaks of higher accuracy at certain 'k' values, like near k=10 and k=30, indicating that these might be more optimal choices for 'k' for the specific classification task. In the context of the investigation, it was decided at a k value of 13 was optimal to increase model performance while decreasing the potential for overfitting.

Conclusion

Summary

For question 1, as we anticipated, we found that attractiveness is more important to males than it is to females when deciding if they want a second date. We also found that the most significant trait overall for both males and females was fun. This was different from what we originally expected for males. However, we did anticipate that fun would be the most significant trait for females. When answering question 3 it was also evident that predictive models can be used to determine if an individual has a chance of a second date occurring based on the ratings in the dataset.

Critique of Methodology

When revisiting our methodology after completing our project, we identified a few weaknesses, with a primary focus on the need for a more diverse range of data analysis techniques. For instance, integrating hypothesis testing could have provided additional perspectives and analytical angles to more comprehensively address

our research questions. In hindsight, we acknowledge that diversifying our analytical approaches would have strengthened the credibility of our conclusion. Another area where we identified a flaw in our methodology was the inclusion of certain analytical methods that did not significantly contribute to the overall conclusion such as finding the average rating for each trait. Although this gave us an idea of what the most important trait might be, it did not give us a very quantifiable answer.

Reliability/ Limitations

Some limitations of the data set include, the dataset primarily consists of participants from Columbia's graduate and professional schools, which may not be representative of the broader population, the sample size is relatively small, the dataset only includes information from the first date in each session, and the brief nature of the speed dates (four minutes each) may not accurately reflect the dynamics of longer, more natural interactions. A small sample size with 276 observations may not capture the full spectrum of experiences and preferences. A larger sample size would provide more robust and reliable results. Only including data from one date overlooks potential variations and trends that might emerge over multiple interactions, missing the chance to analyze the evolution of relationships. Four minutes is a very short amount of time to get to know someone; participants most likely did not have time to form a genuine connection or assess each other thoroughly.

Appropriateness of the statistical analysis

The variables in the dataset are categorical, making linear regression models appropriate when determining our research questions. Finding the mean rating of each trait and looking at the distributions for each trait was also necessary when determining which trait was most significant.

Future Research

For some future research that could be done using our results, one could be finding if the trait that is most significant differs between races. Another thing you could research is looking at if age impacts the trait that is most significant.

References

Andrew, Andrew, Rushton, J., Cade, B., Ungil, C., Hennig, C., Lakeland, D., Anonymous, Harlan, Anonymous, W., Jd, Val, Alper, P., Anoneuoid, Lehman, D., Anon, Eckles, D., Joshua, ... Steven. (2008, January 21). Statistical Modeling, causal inference, and social science. Statistical Modeling Causal Inference and Social Science. https://statmodeling.stat.columbia.edu/2008/01/21/the_speeddating_1/

Chatgpt. (n.d.). https://chat.openai.com/