# A Brief Analysis of Speed Dating

Group 003: Tim Brown, Hannah Striebel, Derien Weatherspoon

22 April 2021

## Introduction

Over the course of this project, we hope to gain insight into how people view potential significant others through the use of a speed dating dataset collected from students at Columbia University. This dataset consists of 276 "dates," where a male and female pair had a four minute speed date and then filled out a form rating various attributes about their date from 1 to 10. The dataset also includes information about each partner, such as their age, race, and gender.

```
Rows: 276
Columns: 22
$ DecisionM        <int> 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1,...
$ DecisionF        <int> 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0,...
$ LikeM            <dbl> 6, 8, 10, 9, 7, 6, 2, 7, 8, 5, 3, 7, 8, 4, 6, 7, 8...
$ LikeF            <dbl> 7, 7, 6, 7, 5, 6, 6, 6, 7, 8, 7, 8, 8, 3, 6, 7, 8,...
$ PartnerYesM      <int> 5, 4, 10, 7, 8, 6, 1, 7, 8, 5, 1, NA, 5, 1, 5, 7, ...
$ PartnerYesF      <int> 5, 3, 2, 8, 5, 1, 5, 6, 10, 7, 7, 5, 1, 2, 6, 5, 5...
$ AgeM             <int> 27, 22, 22, 23, 24, 25, 30, 27, 28, 24, 25, 30, 23...
$ AgeF             <int> 21, 22, 23, 24, 26, 26, 21, 23, 25, 25, 25, 23, 23...
$ RaceM            <fct> Caucasian, Caucasian, Asian, Caucasian, Latino, Ca...
$ RaceF            <fct> Caucasian, Asian, Asian, Caucasian, Other, Caucasi...
$ AttractiveM      <dbl> 6, 7, 10, 9, 7, NA, 3, 6, 8, 5, 5, 10, 6, 4, 6, 7,...
$ AttractiveF      <dbl> 5, 6, 4, 7, 5, 5, 7, 5, 8, 8, 6, 8, 8, 4, 6, 6, 8,...
$ SincereM         <int> 8, 9, 10, 9, 9, 8, 6, 7, 7, 6, 6, 10, 10, 6, 7, 8,...
$ SincereF         <int> 8, 6, 7, 9, 10, 10, 6, 9, 8, 7, 8, 9, 8, 7, 8, 7, ...
$ IntelligentM     <dbl> 8, 10, 10, 9, 8, 7, 7, 8, 8, 8, 6, 10, 9, 7, 6, 7,...
$ IntelligentF     <int> 8, 8, 9, 7, 8, 6, 7, 9, 5, 9, 9, 8, 10, 8, 8, 7, 8...
$ FunM             <int> 8, 6, 10, 9, 9, 7, 5, 6, 10, 5, 5, 10, 8, 3, 6, 6,...
$ FunF             <int> 2, 4, 4, 6, 8, 4, 4, 5, 9, 7, 7, 5, 7, 2, 7, 6, NA...
$ AmbitiousM       <int> 7, 6, 10, 9, 7, 7, 8, 6, 8, 8, 7, 10, 10, 6, 8, 6,...
$ AmbitiousF       <int> 2, 9, 3, 5, 5, 6, 6, 9, 8, 9, 7, 9, 7, 6, 7, 7, 8,...
$ SharedInterestsM <dbl> 6, 5, 10, 9, 7, 7, 7, 5, 10, 6, 3, NA, NA, 4, 5, 6...
$ SharedInterestsF <dbl> 2, 4, 2, 7, 5, NA, 7, 7, 9, 7, 8, 5, 7, 2, 6, 5, 7...
```

By leveraging the information in this dataset, we hope to find answers to a number of questions pertaining to dating. Do people prefer to date inside or outside their own ethnicity? Are males or females more likely to want a second date? Do females tend to rate males higher than males rate females? In what categories? Which features (attractiveness, sincerity, intelligence, etc) are most predictive of wanting a second date? Does this differ by gender? Age? Race?

By answering these questions, we hope to gain quantitative insight and understanding into the complex human experience of dating.

## Methods

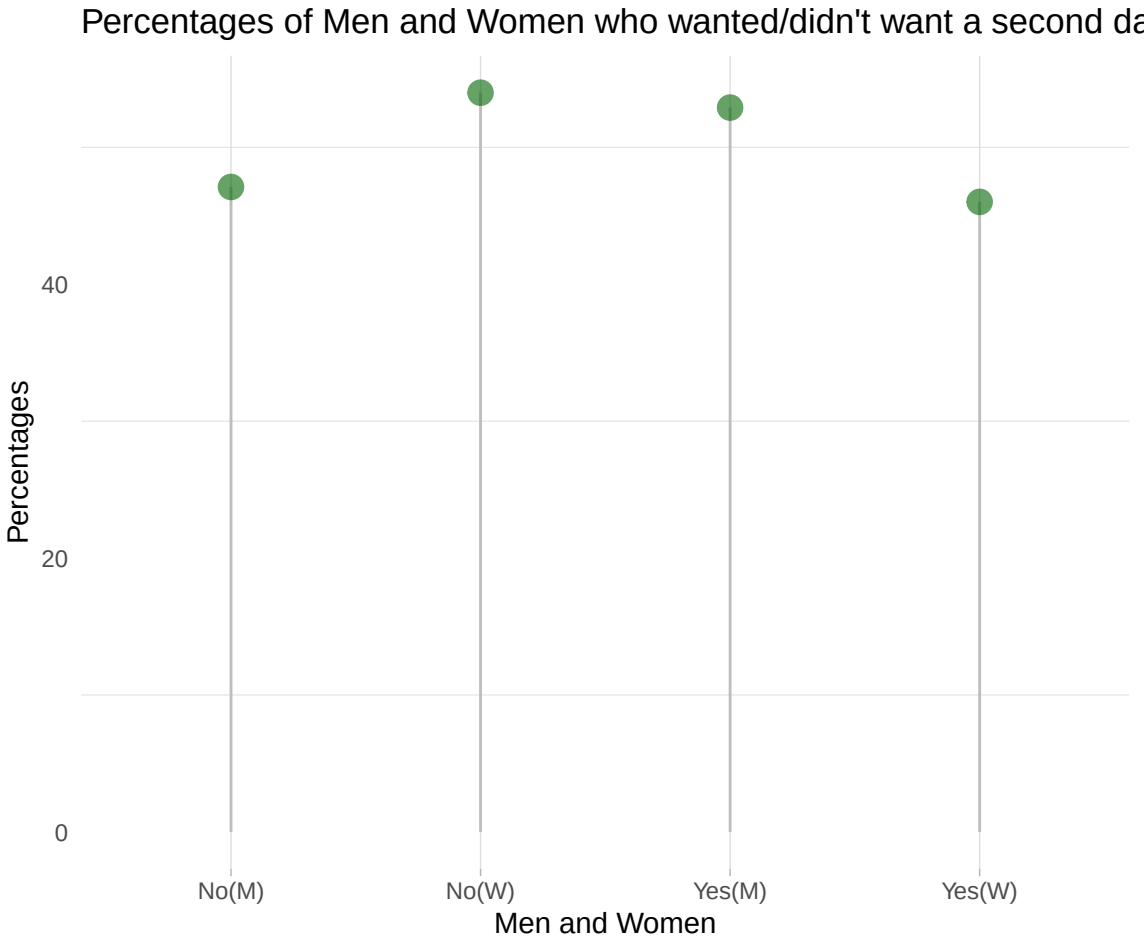
In order to analyze gender alongside ethnicity, the dataset will be filtered by gender and grouped by ethnicity. Analysis can then be done by plotting the data on a bar chart to observe any differences. Analysis of whether or not men or women were more likely to want a second date can be made by comparing the percentage distributions on a singular pie chart. Furthermore, the estimation for a chance at another date men and women gave each other can be found by plotting the ratings each gender gave their date on two different bar plots.

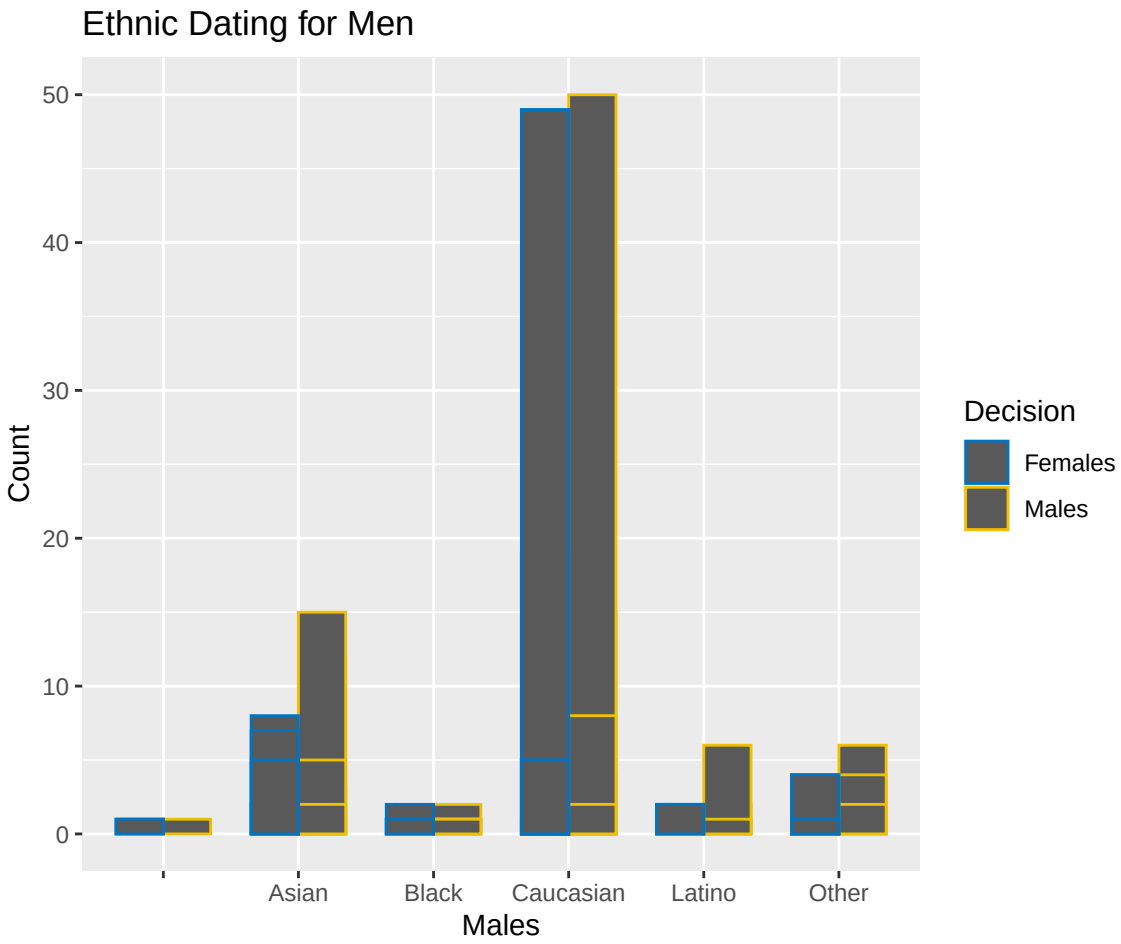
To analyze the potential differences between the genders, very simple measures can be taken. Each factor is already separated by gender so the comparison for any individual factor, such as Fun or Sincerity, can be made by directly comparing the distribution of numerical ratings on a histogram, as well as comparing the averages of those distributions. To visualize some of the surface-level disparities directly related to race, filtering can be used to sort each group on whether or not the race of their partner matched their own race.

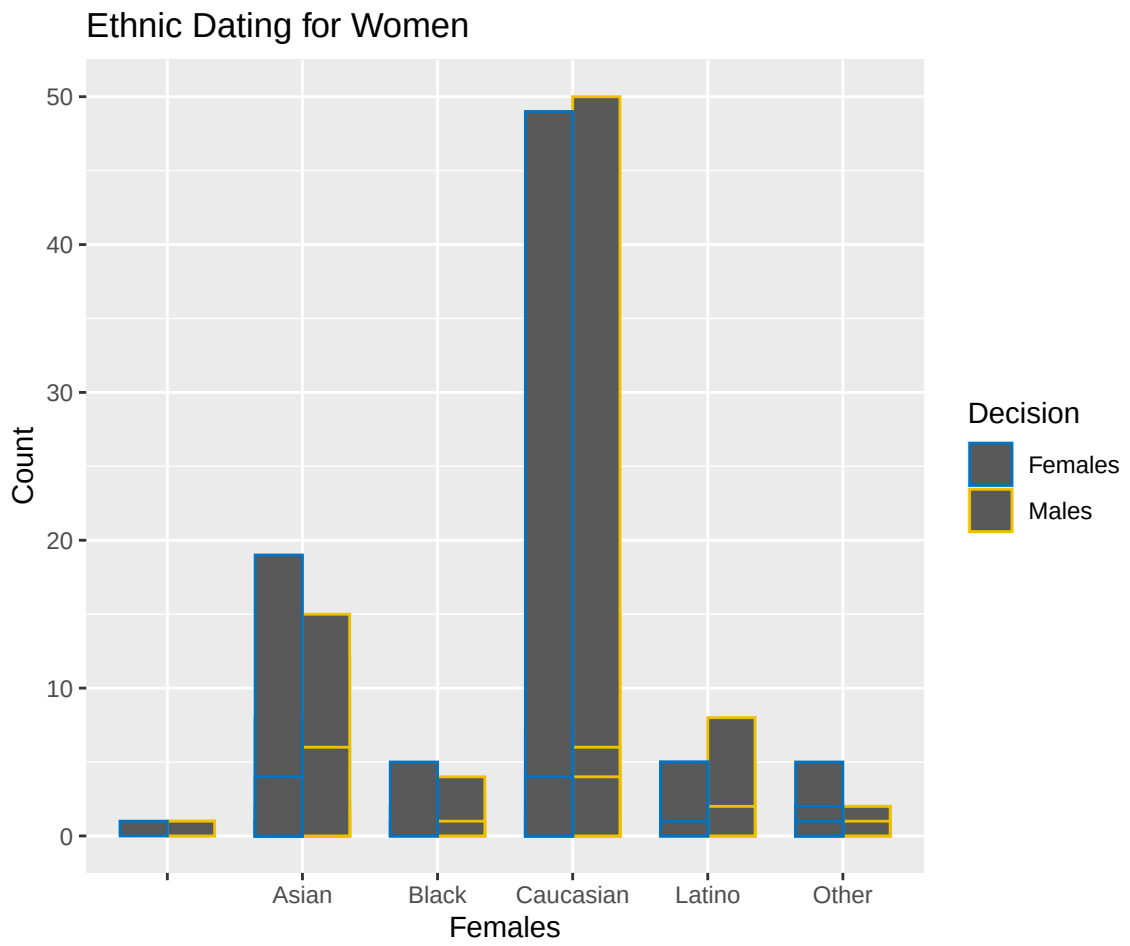
To determine which features are most predictive of wanting a second date, the ratings of features will be plotted, separated by whether or not the individual wanted a second date. If the distribution of ratings for a given feature for both "yes" and "no" are similar, then that feature is not predictive of wanting a second date. However, if the distribution of ratings for "yes" is significantly higher or lower than that of "no", this indicates that this feature is predictive of wanting a second date. Multiple Linear Regression can also be used to attempt to determine which variables are most predictive.

## Results

Men were more likely to want another date after the first date more than women were. The difference in likeliness was minute, showing a modest difference of 7% between the two. Men were also more likely to estimate that their date would be interested in a second date, with their ratings clustered around the top. On the other hand, women's estimations were clustered around the middle. When looking at the results of ethnicity with gender, there were not any great disparities to be found from it. Men and women both went on dates with their own ethnicity and from outside their ethnicity at about an equal amount. No gender or ethnicity had a tendency to date specifically only one ethnicity, however, ethnicity to play a role in influencing for some men and women. For instance, Asian men and women are more likely to date within their own ethnicity.

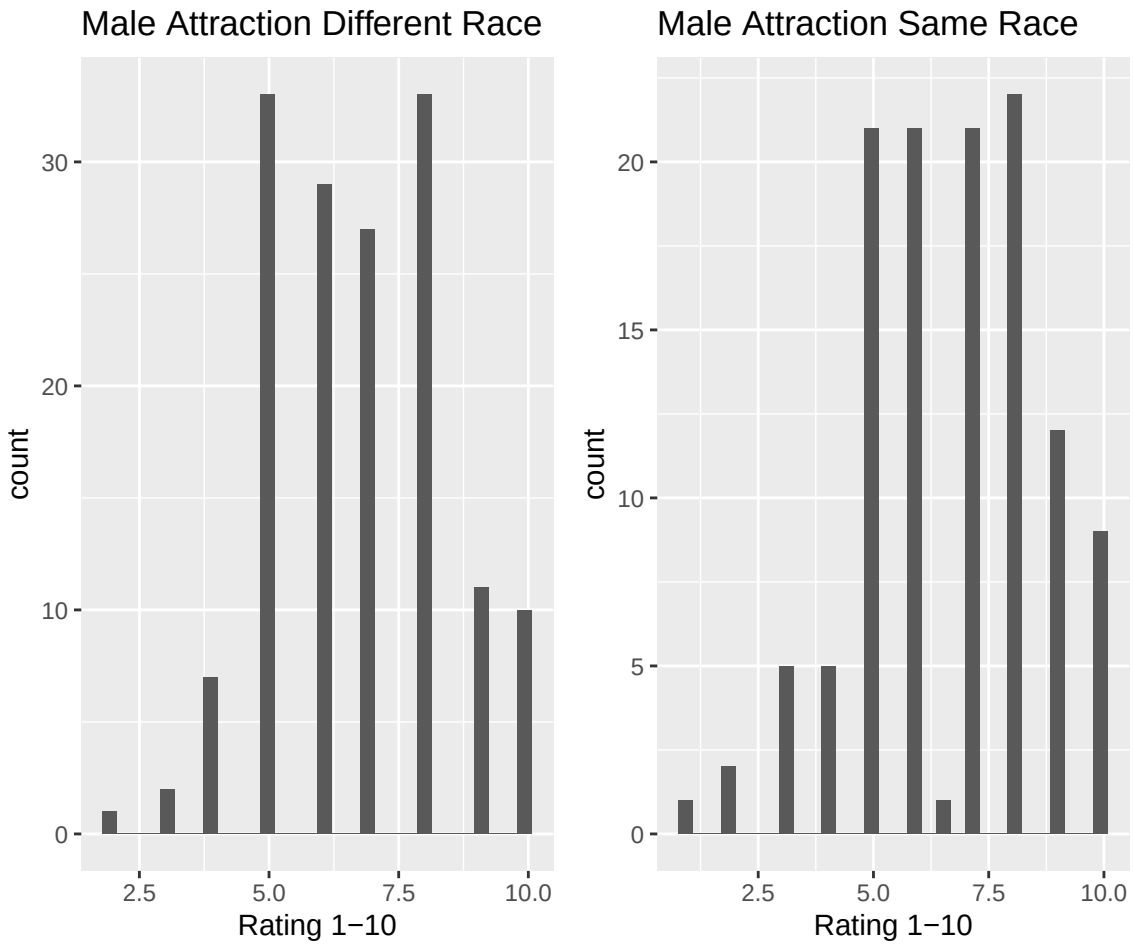

Percentages of Men and Women who wanted/didn't want a second da
Percentages
40
20
0
No(M)
No(W)
Yes(M)
Yes(W)
Men and Women


Ethnic Dating for Men
Count
50
40
30
20
10
0
Asian
Black
Caucasian
Latino
Other
Males
Decision
Females
Males




The distribution of ethnicity dating decisions are for the most part, evenly spread out throughout the sample of men and women.

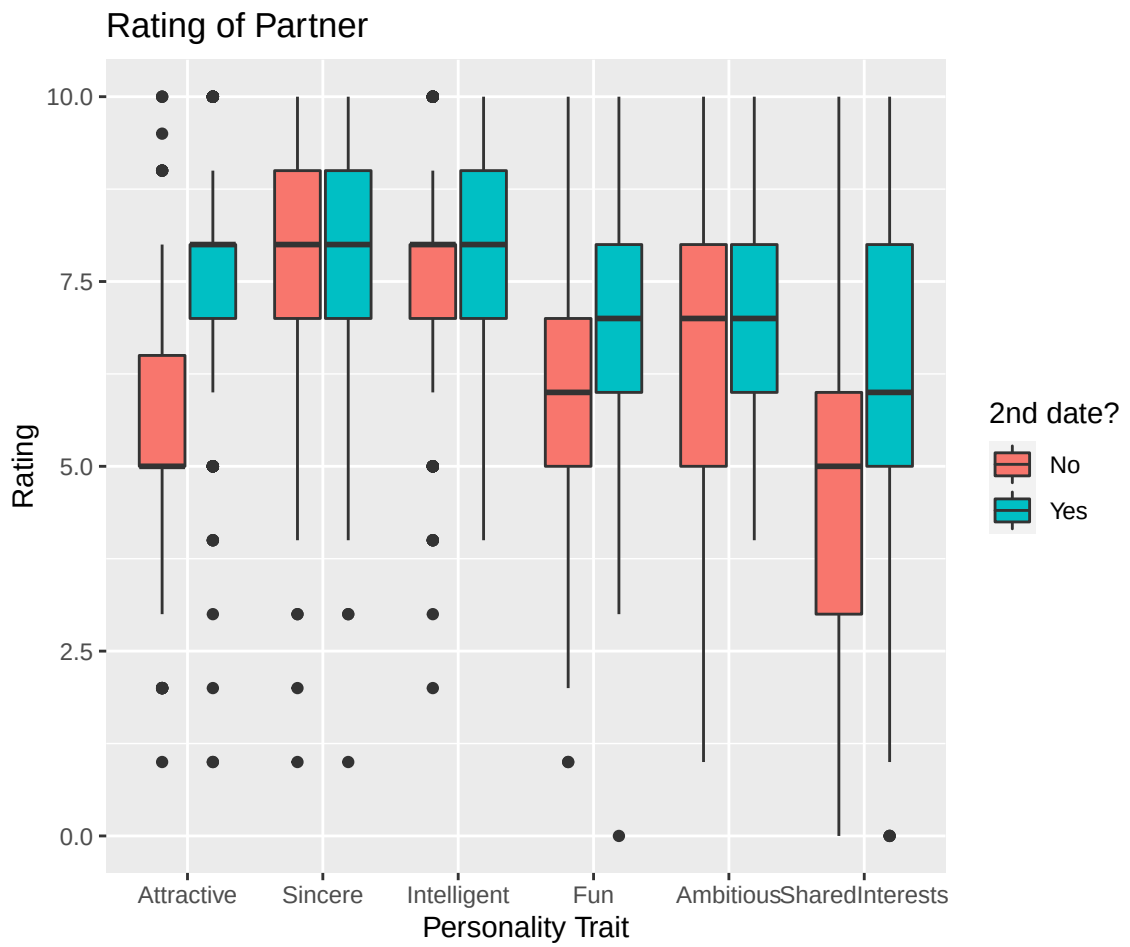
Most category ratings displayed relative parity between the two sexes, though some distinctions arose upon further analysis. Men, on average, rated women's Shared Interests higher than women rated men's. In both the Sincerity and Fun categories, women also rated men on a more even distribution than their male counterparts who clustered their ratings more densely around 6,7 and 8; though this was very subtle. When race became a factor and the two sexes were separated into two groups for analysis, Same Race and Different Race. The distribution for ratings given by women nearly didn't change at all. However, there was a subtle drop in attractiveness ratings given by men when their partner was not of the same race as they were.




Though given the sample size, this may be due to a few outliers as it is a subtle shift in the distribution yet the means are similar.

Which features are most predictive of wanting a second date? Does this differ by gender? Age? Race?

As you can see from the box-plots below, the feature for which the "Yes" and "No" distribution are most dissimilar, and thus most predictive of wanting a second date is `Attractive`. In addition, `Fun` and `Shared Interests` also show a gap between "Yes" and "No."




When the dataset is divided by gender, the distribution of ratings is quite similar to the overall distribution.

When considering race, Caucasian ratings match the overall ratings, most likely because it is the most represented group (1854 samples out of 3276). For Asians, the most predictive feature is also `Attractive` followed by `Fun` and `Shared Interests`, but with the interesting caveat that the overall ratings in all categories are lower. For Blacks, `Shared Interests` seems to be most predictive, with `Attractive` in second. For Latinos, there's a pretty big split between 'yes' and 'no' for most of the features, with `Attractive`, `Fun`, and `Intelligent` being the most predictive.

The most interesting results when answering this question was the distinction between age groups. Ages 18 to 24 still value `Attractive` and `Shared Interests`, but this is the only category where a feature, `Ambitious`, is rated higher for 'no' than for 'yes'. This indicates that young adults see being ambitious as a turn off. The age range of 25 to 30 matches the overall distribution. For adults aged 30 to 55 (the max age in the study), as always, `Attractive` and `Shared Interests` show the largest differences, and interestingly enough, `Ambition` also appears to be a predictor.

By creating a linear model using all the features and then deriving the most parsimonious model, we can analytically determine which features are most predictive.

```
# A tibble: 3 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      -0.423     0.0752     -5.62 3.37e- 8
2 Attractive        0.104     0.0121      8.58 1.45e-16
3 SharedInterests   0.0446    0.0104      4.30 2.12e- 5
```

```
# A tibble: 1 x 2
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.258         0.254
```

The results from the linear model align with what we previously discovered from the box plots - `Attractive` and `Shared Interests` are the most predictive features of wanting a second date. However, it is worth noting that this only results in an r-squared value of about .25, which means that only a quarter of the variation in the data is accounted for by our linear model. This is most likely due to the fact that each person in the study has individual tastes, and that even though `Attractive` and `Shared Interests` are the most predictive *on average*, some people may value other traits more.

# Conclusion

Overall, we were able to gain insights into a number of different questions surrounding the complexities of dating. We found that both men and women rate each other similarly except in a few select categories. Men also seem to perceive their own race as more attractive. Using our analysis, we learned that more men in general would like to go on dates, more than just a singular time. Although race was shown to not be a huge factor in influencing second dates, it still had some influence. And finally, `Attractive`, `Shared Interests`, and `Fun` are the most predictive features of wanting a second date.

The dataset is from Columbia University, which is a reliable source. The dataset is valid for evaluating interpolated results, but may not extrapolate well. In addition, From the description of the data, the only participants were college students. This could bias our findings and potentially render them inaccurate when applied to the general populace. This study also focusses solely on heterosexual speed dates and further narrows the group that the findings could shed light on. Due to the study focusing on speed dating, which are dates around 4-5 minutes, the data could show bias against people who need more than that time given to give ratings on someone. Furthermore, due to the relatively small sample size of the data, outliers are bound to have a much more pronounced effect on the overall findings in certain test cases. While the methods used throughout the analysis are statistically sound, there was not a concerted effort to clean the data/cull obvious outliers from the dataset before analysis.

In the future, we could improve our analysis through a number of means. Instead of merely graphing, we could perform hypothesis tests to see if differences in groups (gender, for example) were statistically significant from one another. Further analysis could be done using predictions instead of simply visualizing the data and interpreting it. In addition, there are a whole host of questions that could be further explored with this data. Given the time that has elapsed since the dataset was collected, we could also delve into the potential shifts in these categories by examining current college students alongside the general population of today. With this more recent repetition we would be able to observe and document any potential changes in preferences, priorities, desirable traits, as well as open the study up to be both more inclusive and comprehensive.

# Sources

Gelman, A. and Hill, J., *Data analysis using regression and multi-level/hierarchical models*, Cambridge University Press: New York, 2007.

Müller, K., & Wickham, H. (n.d.). *Group by one or more variables — group_by*. Https://Dplyr.Tidyverse.Org. Retrieved April 22, 2021, from https://dplyr.tidyverse.org/reference/group_by.html

Wickham, H., Chang, W., Henry, L., Pedersen, T., Takahashi, K., Wilke, C., ... Dunnington, D. (n.d.). *Function reference*. ggplot2 Reference. https://ggplot2.tidyverse.org/reference/.