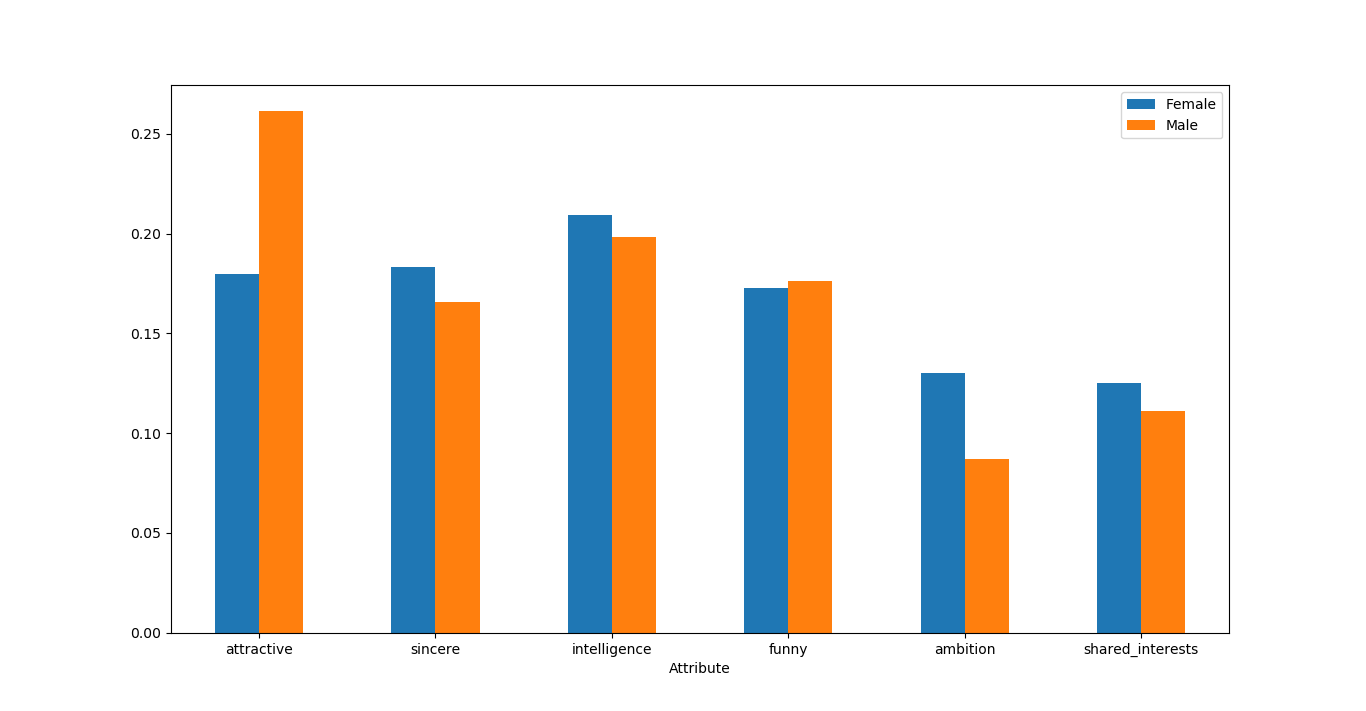
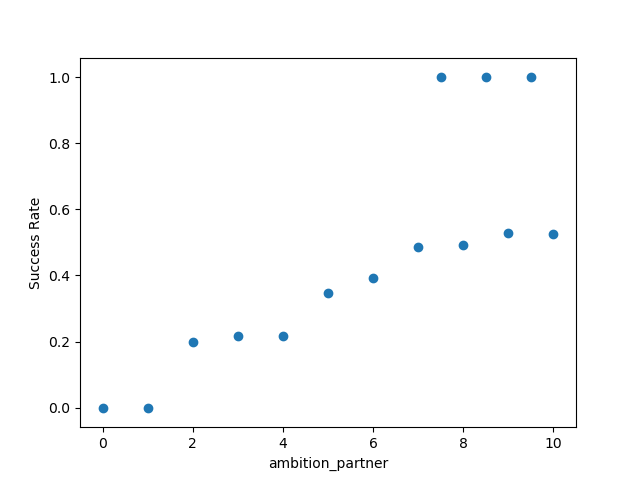
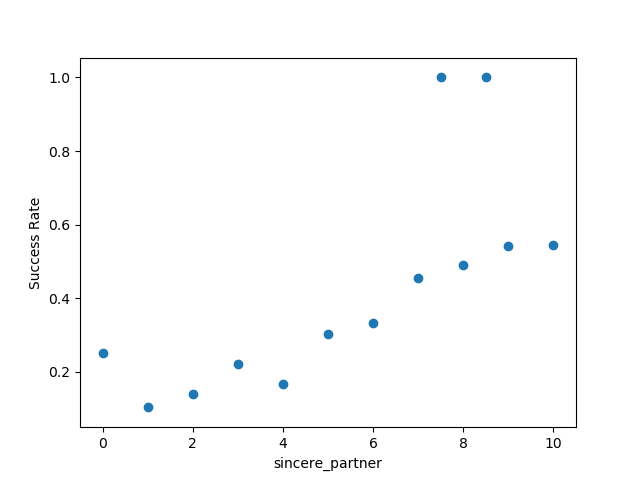
Data Mining Homework 2

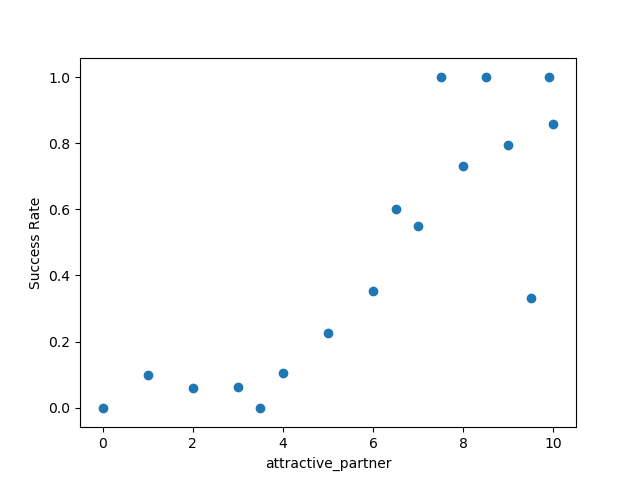
Hrishikesh Viswanath ([hviswan@purdue.edu](mailto:hviswan@purdue.edu))



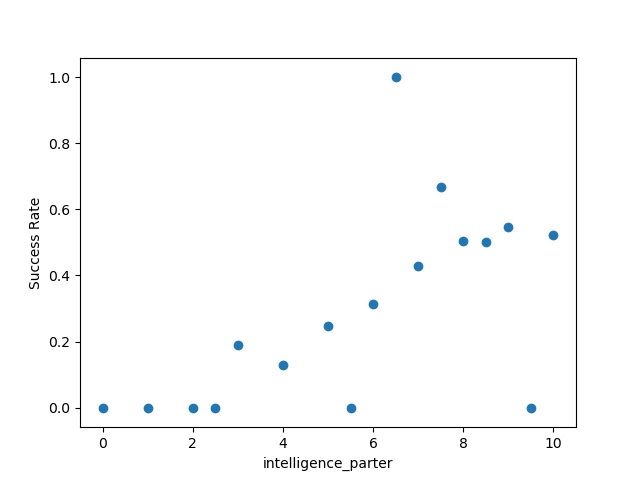
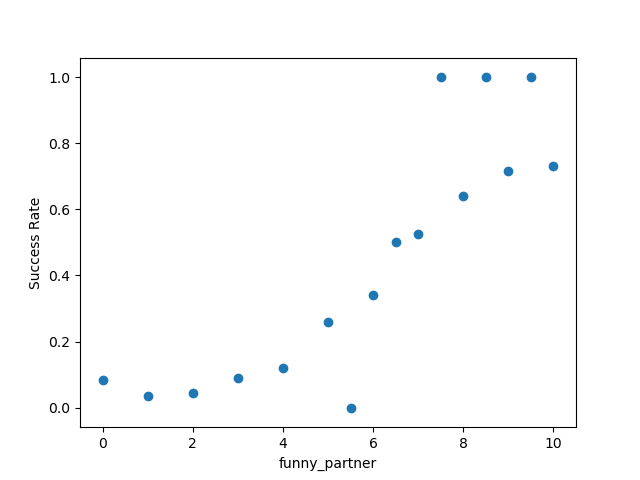
* As per this graph, men give a higher preference to attractive women while women give more importance to ambition than men.
* In other parameters, they are somewhat equal, although women give slightly higher importance.
* Overall, excluding attractiveness, both prefer intelligence first, followed by sincerity and funny nature.

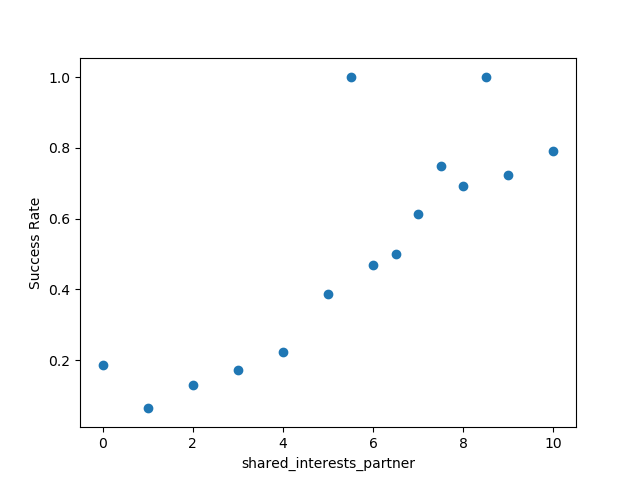




Graphs for sincerity and ambition have a very gradual slope. This implies that they are somewhat important. However, there are some values that are extremely high. These are outliers. A very small percentage of people said yes to second date even when they were insincere or less ambitious. 

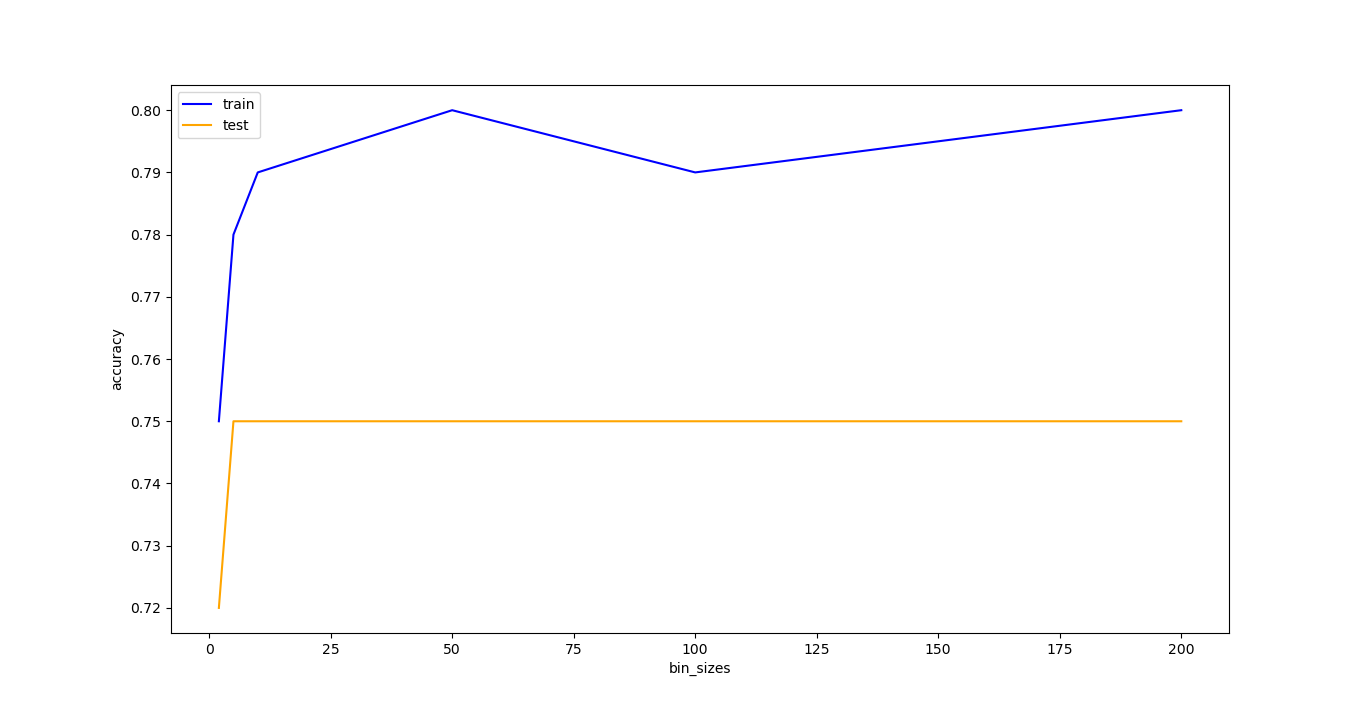
Graphs with higher slope (attractiveness) denote more important qualities. Small increase in those attributes greatly increases chances of a second date

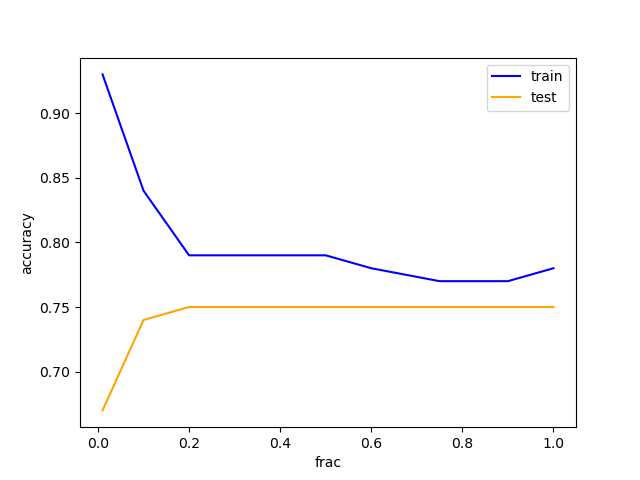
  
Intelligence has a higher threshold than other attributes. People with low intelligence almost have a zero success rate. However, a further increase in intelligence does not greatly vary the success rate



All of the above graphs have positive slope, implying that increasing the rating of the attribute increases the likelihood of the person opting for a second date.

Shared interests graph also has a high slope, which denotes that people greatly value other people who have a lot in common with them.

  
As bin size increases, train accuracy increases. As the number of bins increases, there are fewer values in each bin. However, this implies that it corresponds better with training data. The values in the bin become too constrained and specific to training data, hence training performance increases. However, testing remains the same because testing error doesn’t decrease much as the model fits training data more closely.



Initially, training data is very high and testing data is very low because of overfitting. It perfectly fits the distribution. There are too few elements in the test sample, so the probability values are too specific to train sample and not at all representative of the entire data set.

As training sample size increases, the variability of the data increases, so naïve bayes fails to perform because it is a generative model.