CS57300: Assignment 2

Pavani Guttula
pguttula@purdue.edu

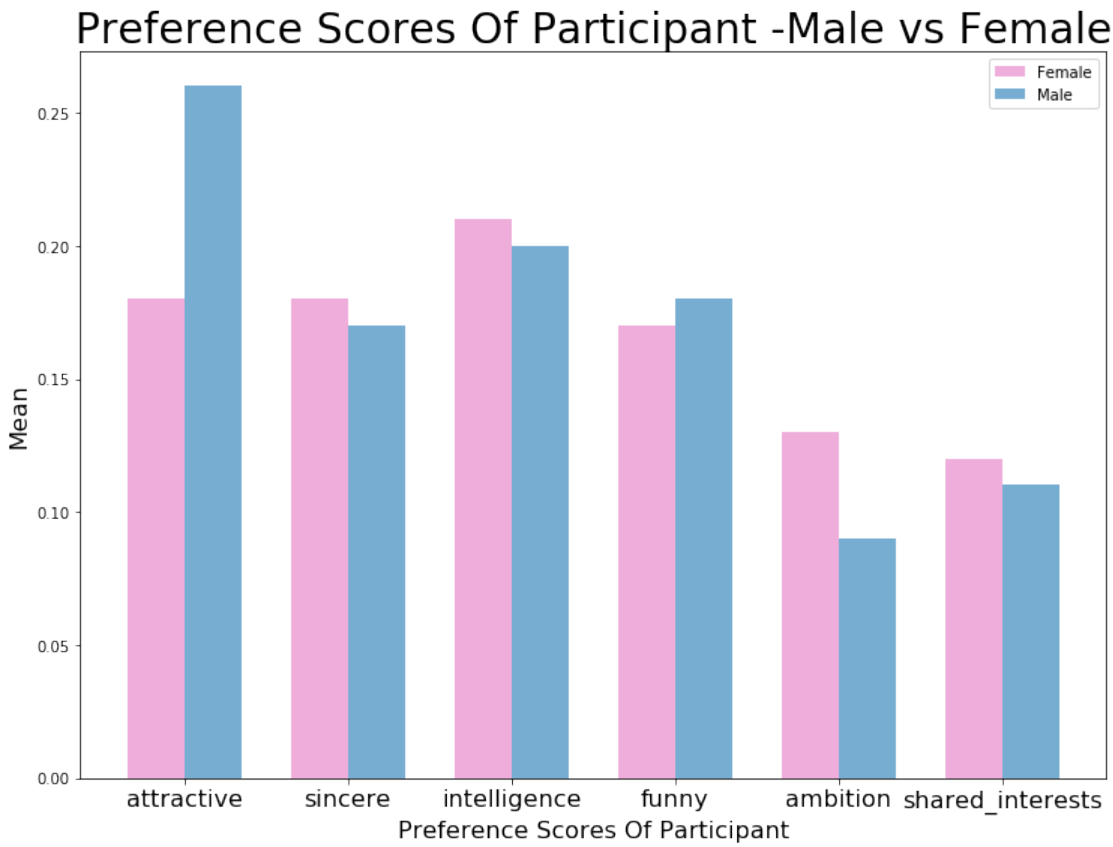
February 13, 2019

1 Preprocessing

```
Output:
Quotes removed from 8316 cells
Standardized 5707 cells to lower case
Value assigned for male in column gender: 1
Value assigned for European/Caucasian-American in column race: 2
Value assigned for Latino/Hispanic American in column race_o: 3
Value assigned for law in column field: 121
Mean of attractive_important: 0.22
Mean of sincere_important: 0.17
Mean of intelligence_important: 0.2
Mean of funny_important: 0.17
Mean of ambition_important: 0.11
Mean of shared_interests_important: 0.12
Mean of pref_o_attractive: 0.22
Mean of pref_o_sincere: 0.17
Mean of pref_o_intelligence: 0.2
Mean of pref_o_funny: 0.17
Mean of pref_o_ambitious: 0.11
Mean of pref_o_shared_interests: 0.12
```

2 Visualization

Output of 2_1.py:

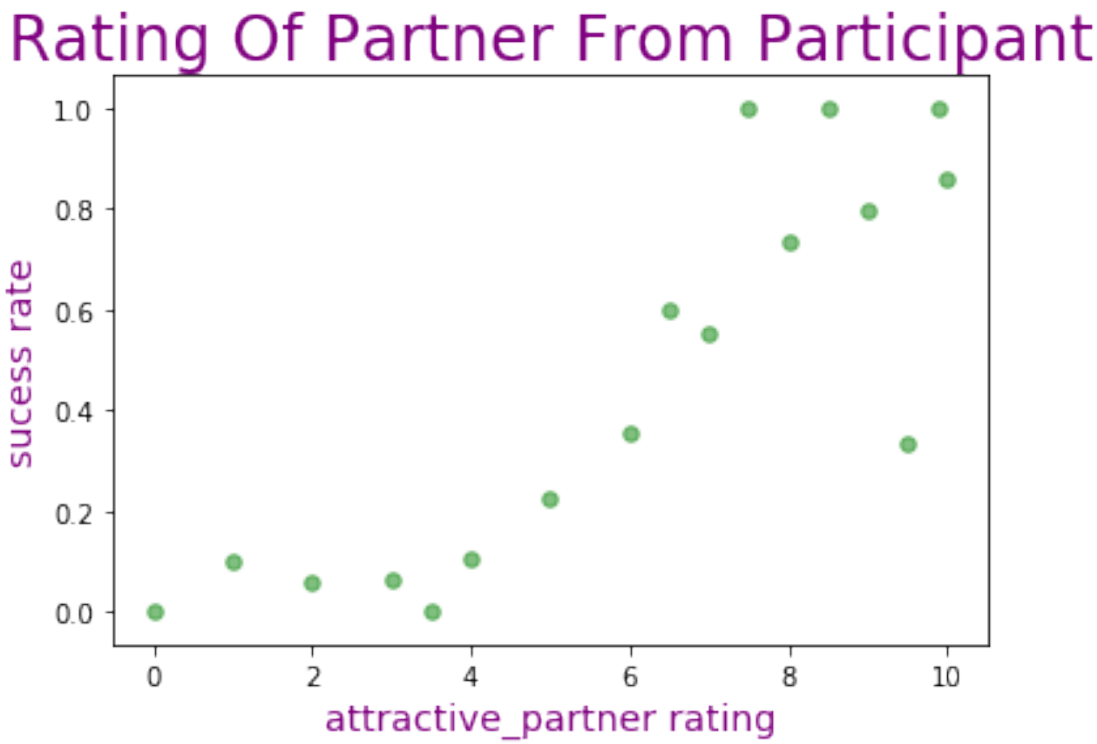


From the barplot, we can see that on an average, the following characteristics are more important to the Males than Females:
Partner is attractive
Partner is funny

And the following characteristics are more important to the females than Males:
Partner is ambitious
Partner is intelligent
Partner is sincere
Partner and the participant has common/shared intrests

We can also observe that there is a huge difference in the male and female prefereneces for the below two characteristics:
Compared to females, for males it is very important that their partner is attractive.
Compared to males, for females it is very important that their partner is ambitious.

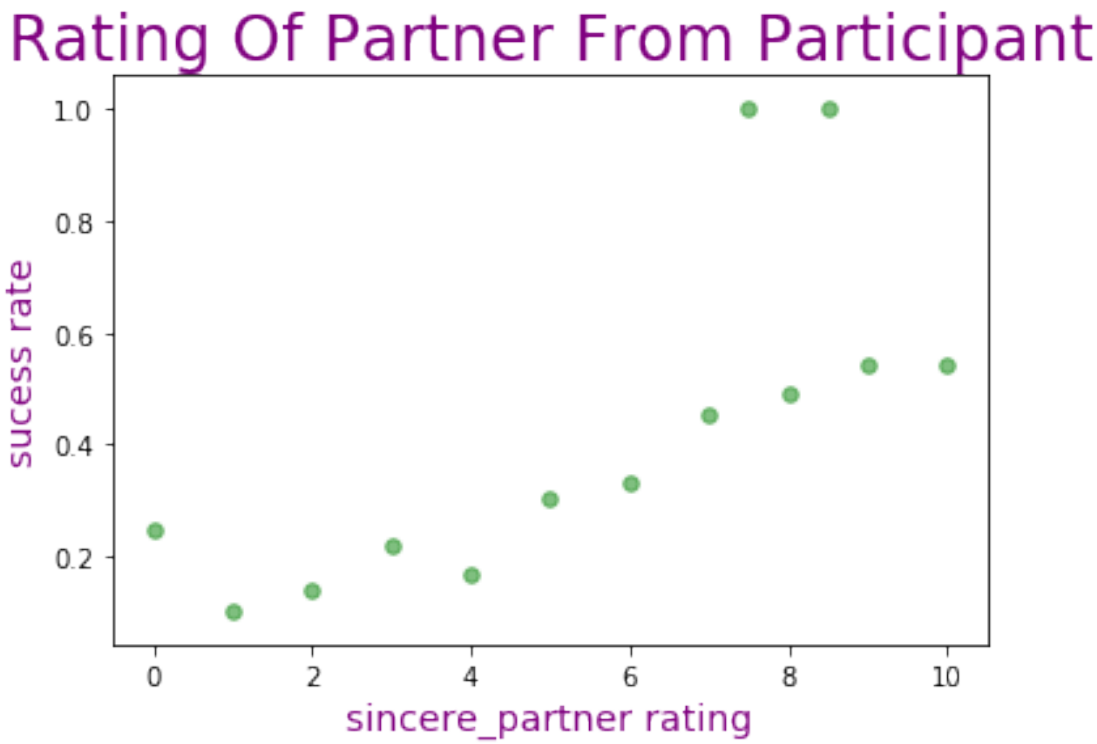
Output of 2_2.py:



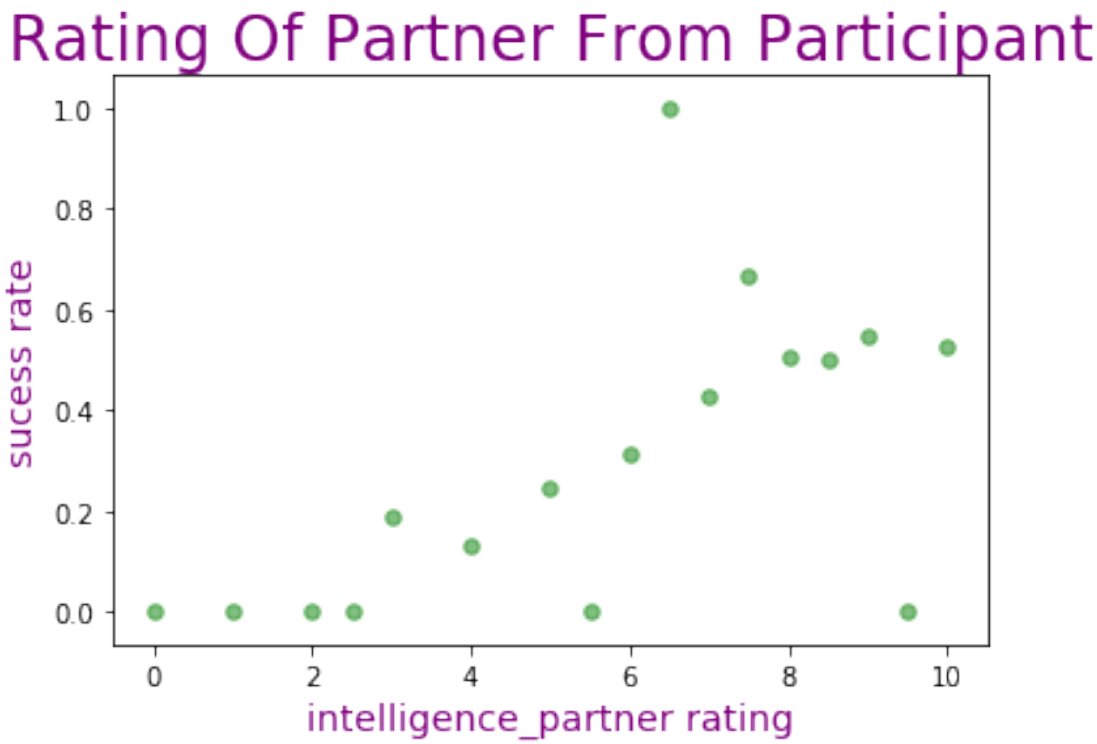
Participants who rate their partner above 6 based on "attractiveness", give their partner a second date more than 50% of times.

Participants who rate their partner above 8 based on "attractiveness", give their partner a second date more than 80% of times.

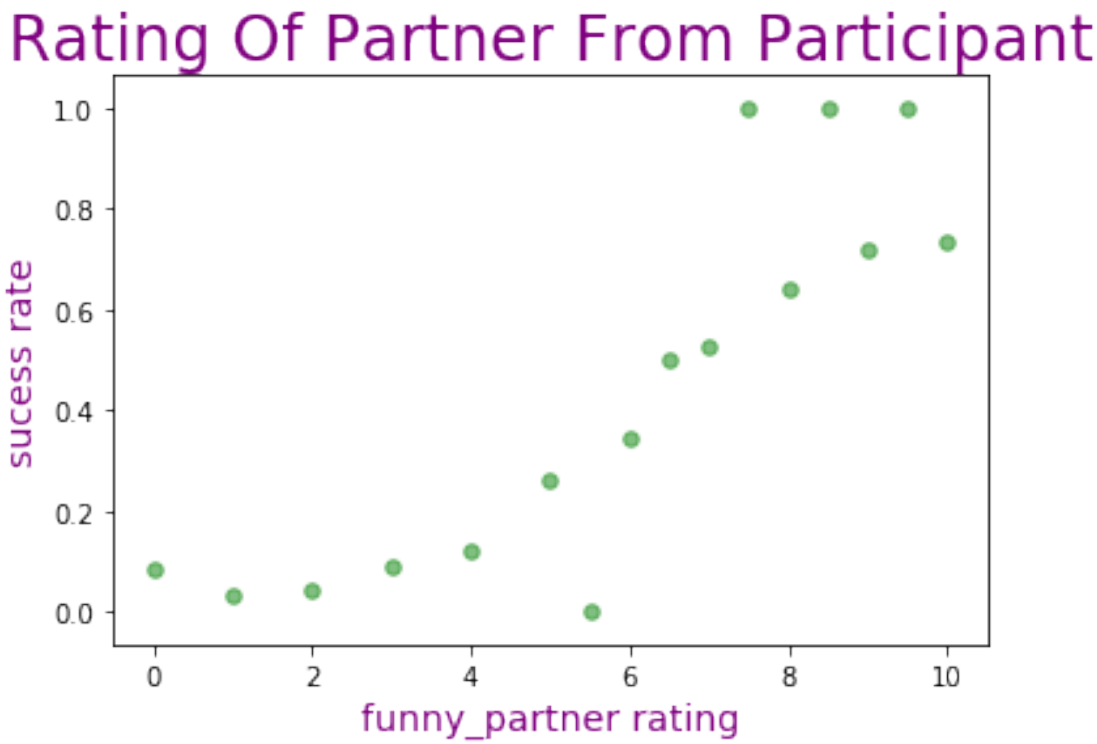
We can see that an increase in the rating shows an increase in the success rate from a rating of 4 and above.



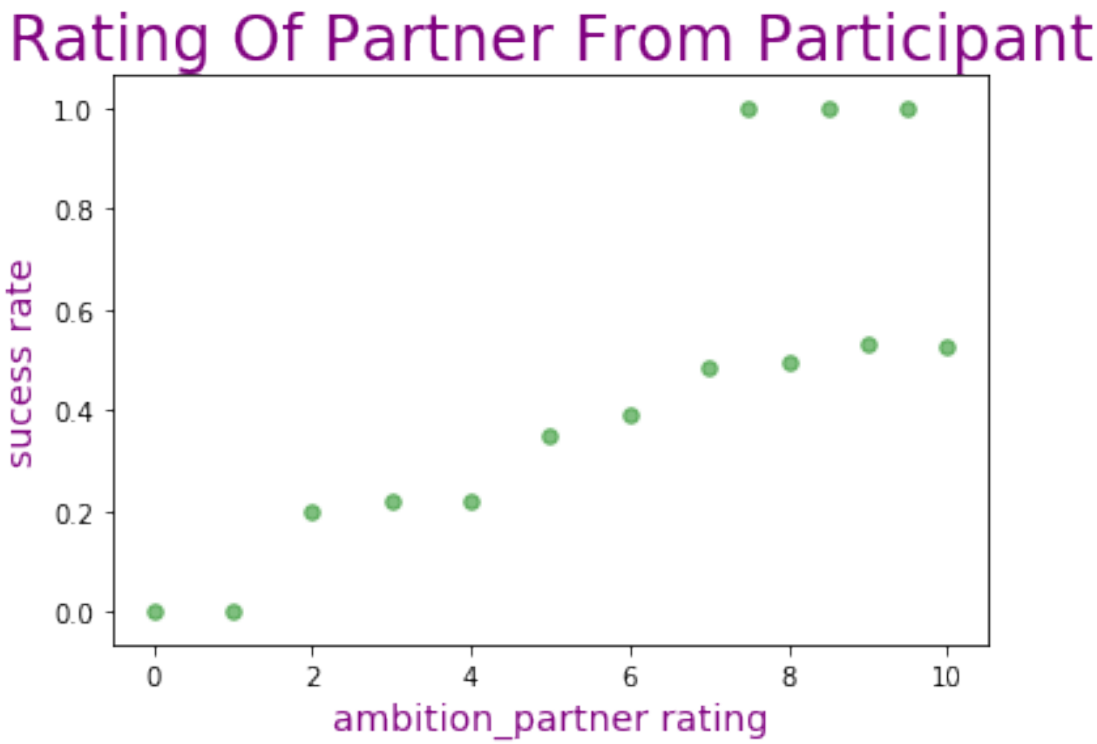
From the above scatter plot, We can observe that a rating above 6 by a participant to the partner based on "sincere" goes on a second date around 50%-60% of the times. We can see that an increase in the rating shows an increase in the success rate overall.



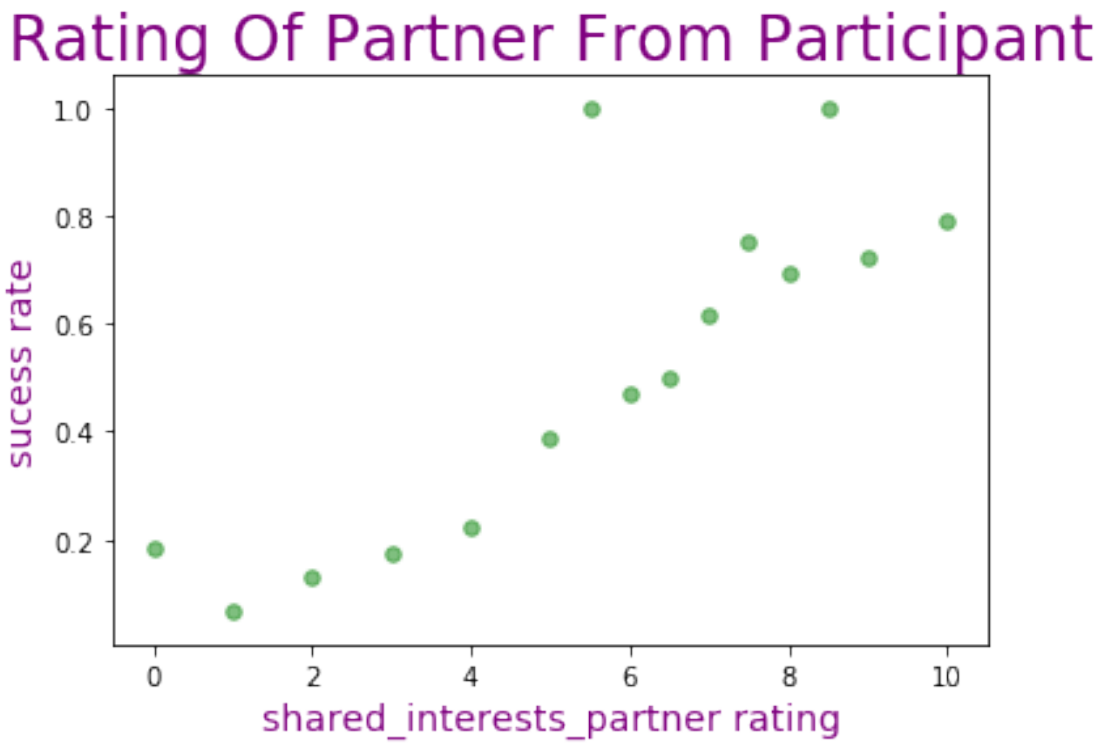
From the abovse scatter plot, we can say that if a particpant gives their partner a rating of 7 or above on "intelligence", it is more(50%-60%) likely that they go on a second date. We can see that an increase in the rating shows an increase in the success rate from a rating of 4 and above.



If a particpant gives their partner a rating of 8 or above on "funny", it is more likely(>60%) that they go on a second date. We can see that an increase in the rating shows an increase in the success rate overall.



If a particpant gives their partner a rating of 7 or above on "ambition", it is more(50%-60%) likely that they go on a second date.We can see that an increase in the rating shows an increase in the success rate from a rating of 4 and above.



We can see that an increase in the rating by the participant on "shared instrests" shows an increase in the success rate overall.

3 Continuous to Catergorical attributes Conversion

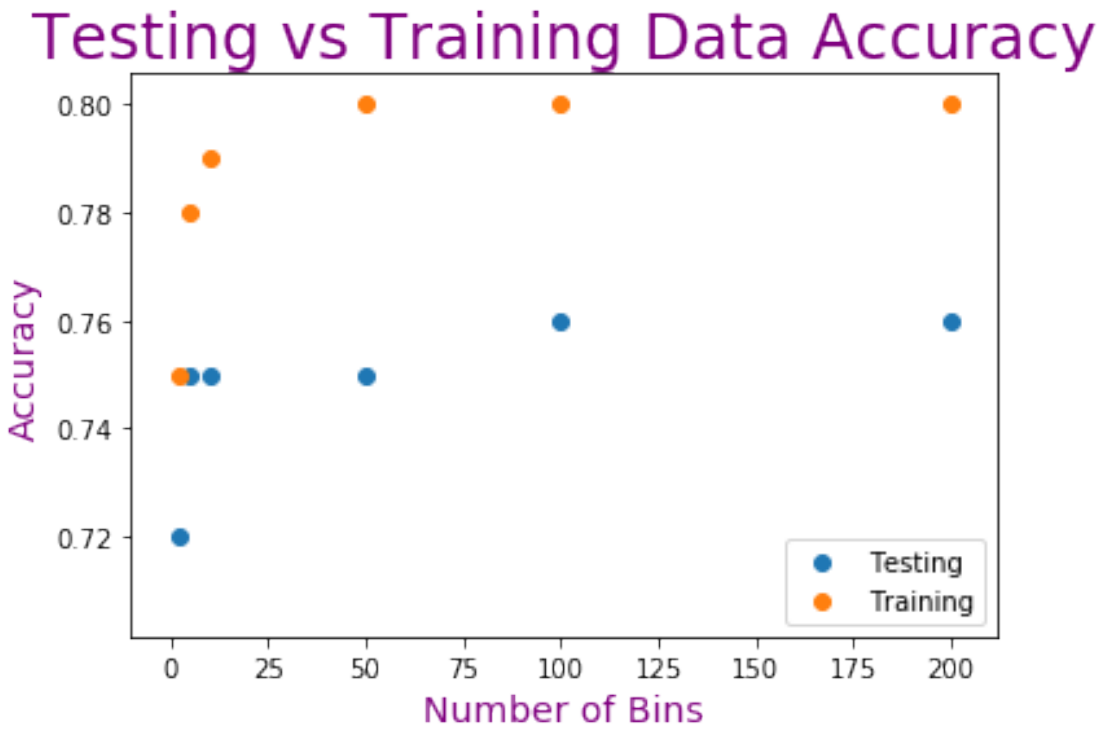
Output of discretize.py:

```
age : [3710, 2932, 97, 5, 0]
age_o : [3704, 2899, 136, 5, 0]
importance_same_race : [2980, 1213, 1013, 977, 561]
importance_same_religion : [3203, 1188, 1110, 742, 501]
pref_o_attractive : [4333, 1987, 344, 51, 29]
pref_o_sincere : [5500, 1225, 19, 0, 0]
pref_o_intelligence : [4601, 2062, 81, 0, 0]
pref_o_funny : [5616, 1103, 25, 0, 0]
pref_o_ambitious : [6656, 88, 0, 0, 0]
pref_o_shared_interests : [6467, 277, 0, 0, 0]
attractive_important : [4323, 2017, 328, 57, 19]
sincere_important : [5495, 1235, 14, 0, 0]
intelligence_important : [4606, 2071, 67, 0, 0]
funny_important : [5588, 1128, 28, 0, 0]
ambition_important : [6644, 100, 0, 0, 0]
shared_interests_important : [6494, 250, 0, 0, 0]
attractive : [4122, 1462, 866, 276, 18]
```

```
sincere : [3392, 2715, 487, 117, 33]
intelligence : [3190, 2286, 1049, 185, 34]
funny : [3313, 3191, 221, 19, 0]
ambition : [2876, 2387, 1070, 327, 84]
attractive_partner : [2418, 2390, 948, 704, 284]
sincere_partner : [3282, 1627, 1388, 353, 94]
intelligence_parter : [3509, 1509, 1497, 193, 36]
funny_partner : [2600, 2296, 836, 733, 279]
ambition_partner : [2804, 2258, 1090, 473, 119]
shared_interests_partner : [2536, 1774, 1269, 701, 464]
sports : [2077, 1687, 1369, 961, 650]
tvsports : [2151, 1383, 1292, 1233, 685]
exercise : [2115, 1775, 1283, 952, 619]
dining : [2797, 2618, 1118, 172, 39]
museums : [2737, 1741, 1417, 732, 117]
art : [2500, 1557, 1517, 946, 224]
hiking : [1855, 1575, 1386, 965, 963]
gaming : [2565, 1522, 1435, 979, 243]
clubbing : [2193, 1668, 1068, 912, 903]
reading : [2827, 2317, 1071, 398, 131]
tv : [1999, 1642, 1216, 1188, 699]
theater : [2300, 1760, 1585, 811, 288]
movies : [2825, 2783, 843, 248, 45]
concerts : [2282, 1752, 1711, 777, 222]
music : [2797, 2583, 1106, 196, 62]
shopping : [1709, 1643, 1201, 1098, 1093]
yoga : [2285, 1392, 1369, 1056, 642]
interests_correlate : [2875, 2520, 758, 573, 18]
expected_happy_with_sd_people : [3292, 1596, 1262, 321, 273]
like : [2560, 2539, 865, 507, 273]
```

4 NBC Implementation

Output of 5_2.py:

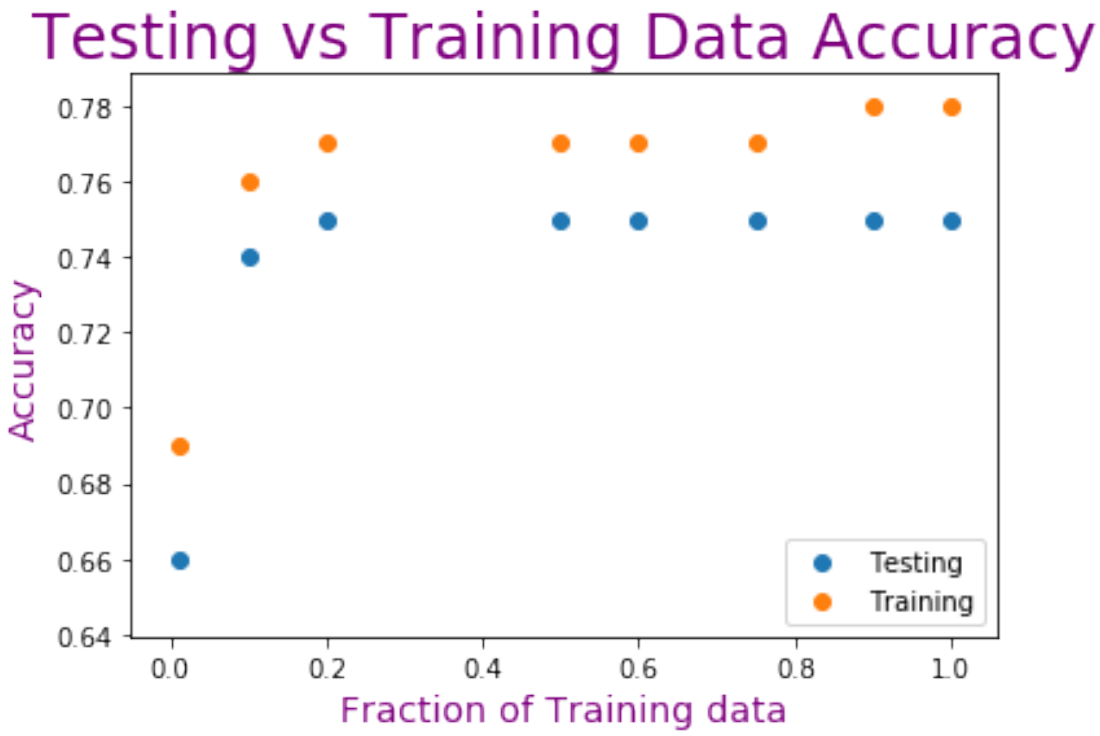


Our NBC model's performace on both training and test data remains almost the same irrespective of the increase in the number of bins from 5.

With an increase in the number of bins from 2 to 5, we do see an increase of accuracy by 3% in both training and test data.

So, Binning the data into 5 or more number would not increase the accuracy of our model.

Output of 5_3.py:



NBC model's performance remains the same on both training and test data with an increase in the fraction of training data after 0.2. So, training on a sample of 0.2 or more fraction of data would not change the accuracy of our model.