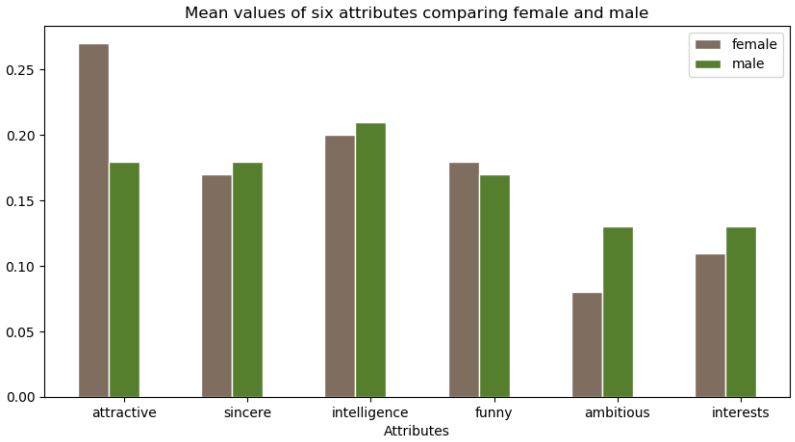
CS573 Assignment 2:

pre_process outputs:

quotes removed from 8316 cells
5707 Standardized field 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_1 output:

Mean of attractive_important in female data: 0.18
Mean of sincere_important in female data: 0.18
Mean of intelligence_important in female data: 0.21
Mean of funny_important in female data: 0.17
Mean of ambition_important in female data: 0.13
Mean of shared_interests_important in female data: 0.12
Mean of attractive_important in male data: 0.26
Mean of sincere_important in male data: 0.17
Mean of intelligence_important in male data: 0.2
Mean of funny_important in male data: 0.18
Mean of ambition_important in male data: 0.09
Mean of shared_interests_important in male data: 0.11
Mean of pref_o_attractive in female data: 0.27
Mean of pref_o_sincere in female data: 0.17
Mean of pref_o_intelligence in female data: 0.2
Mean of pref_o_funny in female data: 0.18
Mean of pref_o_ambitious in female data: 0.08
Mean of pref_o_shared_interests in female data: 0.11
Mean of pref_o_attractive in male data: 0.18
Mean of pref_o_sincere in male data: 0.18
Mean of pref_o_intelligence in male data: 0.21
Mean of pref_o_funny in male data: 0.17
Mean of pref_o_ambitious in male data: 0.13
Mean of pref_o_shared_interests in male data: 0.13



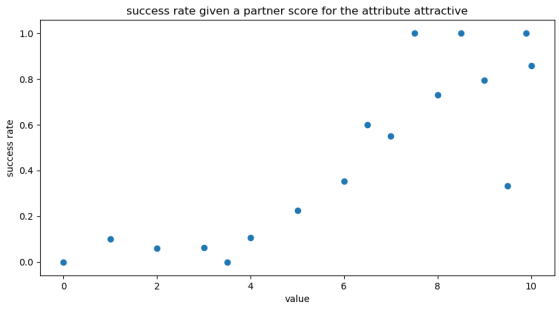
The top three qualities that men prefer are intelligence, sincere, and attractive, while the top three qualities the women prefer are attractive, intelligence, and funny.

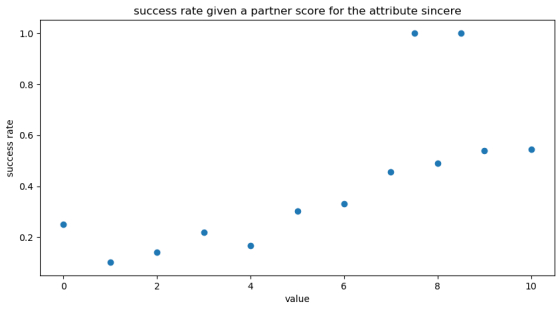
2_2 output:

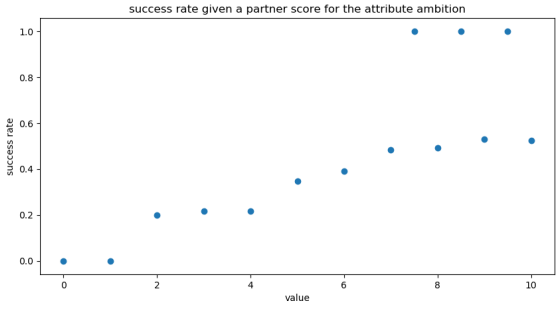
Mean of attractive_important in female data: 0.18
Mean of sincere_important in female data: 0.18
Mean of intelligence_important in female data: 0.21
Mean of funny_important in female data: 0.17
Mean of ambition_important in female data: 0.13
Mean of shared_interests_important in female data: 0.12
Mean of attractive_important in male data: 0.26
Mean of sincere_important in male data: 0.17
Mean of intelligence_important in male data: 0.2
Mean of funny_important in male data: 0.18
Mean of ambition_important in male data: 0.09
Mean of shared_interests_important in male data: 0.11
Mean of pref_o_attractive in female data: 0.27
Mean of pref_o_sincere in female data: 0.17
Mean of pref_o_intelligence in female data: 0.2
Mean of pref_o_funny in female data: 0.18
Mean of pref_o_ambitious in female data: 0.08
Mean of pref_o_shared_interests in female data: 0.11
Mean of pref_o_attractive in male data: 0.18
Mean of pref_o_sincere in male data: 0.18
Mean of pref_o_intelligence in male data: 0.21
Mean of pref_o_funny in male data: 0.17
Mean of pref_o_ambitious in male data: 0.13
Mean of pref_o_shared_interests in male data: 0.13
Number of distinct values for attractive_partner: 17
Number of distinct values for sincere_partner: 13
Number of distinct values for intelligence_partner: 17
Number of distinct values for funny_partner: 16
Number of distinct values for ambition_partner: 14
Number of distinct values for shared_interests_partner: 15
The success rate given 6.0 in attractive_partner is 0.3543022415039769

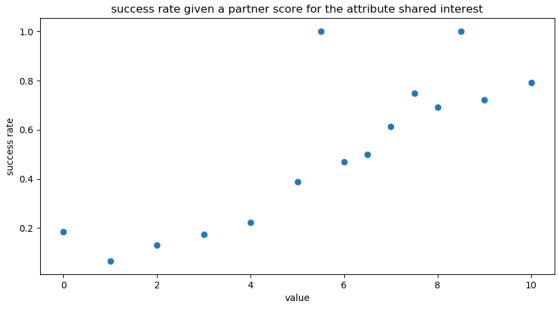
The success rate given 7.0 in attractive_partner is 0.5510355029585798
The success rate given 5.0 in attractive_partner is 0.22608695652173913
The success rate given 4.0 in attractive_partner is 0.10543130990415335
The success rate given 8.0 in attractive_partner is 0.7310679611650486
The success rate given 9.0 in attractive_partner is 0.7955056179775281
The success rate given 3.0 in attractive_partner is 0.06230529595015576
The success rate given 10.0 in attractive_partner is 0.8582677165354331
The success rate given 2.0 in attractive_partner is 0.06091370558375635
The success rate given 1.0 in attractive_partner is 0.1
The success rate given 0.0 in attractive_partner is 0.0
The success rate given 6.5 in attractive_partner is 0.6
The success rate given 7.5 in attractive_partner is 1.0
The success rate given 9.5 in attractive_partner is 0.3333333333333333
The success rate given 8.5 in attractive_partner is 1.0
The success rate given 9.9 in attractive_partner is 1.0
The success rate given 3.5 in attractive_partner is 0.0
The success rate given 9.0 in sincere_partner is 0.54125
The success rate given 8.0 in sincere_partner is 0.49023090586145646
The success rate given 6.0 in sincere_partner is 0.33143399810066476
The success rate given 7.0 in sincere_partner is 0.45540201005025127
The success rate given 5.0 in sincere_partner is 0.30313588850174217
The success rate given 10.0 in sincere_partner is 0.5443686006825939
The success rate given 4.0 in sincere_partner is 0.1680327868852459
The success rate given 3.0 in sincere_partner is 0.22018348623853212
The success rate given 2.0 in sincere_partner is 0.14035087719298245
The success rate given 1.0 in sincere_partner is 0.10344827586206896
The success rate given 0.0 in sincere_partner is 0.25
The success rate given 8.5 in sincere_partner is 1.0
The success rate given 7.5 in sincere_partner is 1.0
The success rate given 7.0 in intelligence_partner is 0.42857142857142855
The success rate given 8.0 in intelligence_partner is 0.5027716186252772
The success rate given 6.0 in intelligence_partner is 0.313824419778002
The success rate given 9.0 in intelligence_partner is 0.5459401709401709
The success rate given 10.0 in intelligence_partner is 0.5232974910394266
The success rate given 5.0 in intelligence_partner is 0.24564796905222436
The success rate given 4.0 in intelligence_partner is 0.12949640287769784
The success rate given 3.0 in intelligence_partner is 0.18867924528301888
The success rate given 2.0 in intelligence_partner is 0.0
The success rate given 1.0 in intelligence_partner is 0.0
The success rate given 0.0 in intelligence_partner is 0.0
The success rate given 6.5 in intelligence_partner is 1.0
The success rate given 8.5 in intelligence_partner is 0.5
The success rate given 7.5 in intelligence_partner is 0.6666666666666666
The success rate given 9.5 in intelligence_partner is 0.0
The success rate given 2.5 in intelligence_partner is 0.0
The success rate given 5.5 in intelligence_partner is 0.0
The success rate given 7.0 in funny partner is 0.5261648745519714
The success rate given 8.0 in funny partner is 0.6394671107410491
The success rate given 4.0 in funny partner is 0.12024048096192384
The success rate given 9.0 in funny partner is 0.7173489278752436
The success rate given 6.0 in funny partner is 0.3421450151057402
The success rate given 3.0 in funny partner is 0.08974358974358974
The success rate given 5.0 in funny partner is 0.26055612770339853
The success rate given 10.0 in funny partner is 0.7320872274143302

The success rate given 1.0 in funny partner is 0.03409090909090909
The success rate given 2.0 in funny partner is 0.0446927374301676
The success rate given 0.0 in funny partner is 0.08333333333333333
The success rate given 5.5 in funny partner is 0.0
The success rate given 6.5 in funny partner is 0.5
The success rate given 9.5 in funny partner is 1.0
The success rate given 7.5 in funny partner is 1.0
The success rate given 8.5 in funny partner is 1.0
The success rate given 6.0 in ambition partner is 0.3921259842519685
The success rate given 5.0 in ambition partner is 0.347165991902834
The success rate given 8.0 in ambition partner is 0.4930232558139535
The success rate given 10.0 in ambition partner is 0.5250596658711217
The success rate given 9.0 in ambition partner is 0.5291479820627802
The success rate given 3.0 in ambition partner is 0.21710526315789475
The success rate given 7.0 in ambition partner is 0.48544973544973546
The success rate given 4.0 in ambition partner is 0.21806853582554517
The success rate given 2.0 in ambition partner is 0.2
The success rate given 1.0 in ambition partner is 0.0
The success rate given 0.0 in ambition partner is 0.0
The success rate given 9.5 in ambition partner is 1.0
The success rate given 7.5 in ambition partner is 1.0
The success rate given 8.5 in ambition partner is 1.0
The success rate given 5.0 in shared interest partner is 0.3868131868131868
The success rate given 6.0 in shared interest partner is 0.46923076923076923
The success rate given 8.0 in shared interest partner is 0.6929698708751794
The success rate given 4.0 in shared interest partner is 0.22344827586206897
The success rate given 7.0 in shared interest partner is 0.6143790849673203
The success rate given 3.0 in shared interest partner is 0.17279411764705882
The success rate given 2.0 in shared interest partner is 0.13111111111111112
The success rate given 9.0 in shared interest partner is 0.7228070175438597
The success rate given 10.0 in shared interest partner is 0.7909604519774012
The success rate given 1.0 in shared interest partner is 0.06598984771573604
The success rate given 0.0 in shared interest partner is 0.18518518518518517
The success rate given 7.5 in shared interest partner is 0.75
The success rate given 6.5 in shared interest partner is 0.5
The success rate given 8.5 in shared interest partner is 1.0
The success rate given 5.5 in shared interest partner is 1.









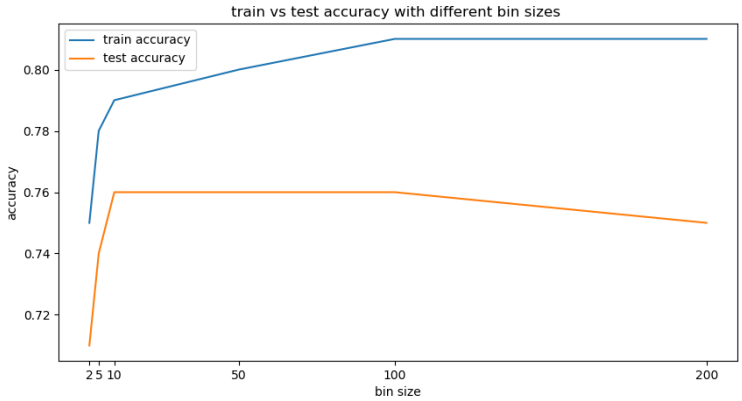
I observe from the above plot that the higher the value each, no matter which attribute, the higher the success rate is.

Output for discrete.py:

age: [3032, 3390, 297, 20, 5]
age_0: [2990, 3362, 371, 16, 5]
importance_same_race: [2980, 2980, 2980, 1213, 1213]
importance_same_religion: [3203, 3203, 3203, 1188, 1188]
pref_o_attractive: [4333, 29, 344, 51, 29]
pref_o_sincere: [1416, 4378, 865, 79, 6]
pref_o_intelligence: [666, 3935, 1873, 189, 81]
pref_o_funny: [1255, 4361, 1048, 55, 25]
pref_o_ambitious: [1963, 2352, 2365, 42, 22]
pref_o_shared_interests: [1506, 2068, 1981, 1042, 147]
attractive_important: [4323, 19, 328, 57, 19]
sincere_important: [546, 2954, 2782, 377, 85]
intelligence_important: [630, 3976, 1861, 210, 67]
funny_important: [1282, 4306, 1070, 58, 28]
ambition_important: [1913, 2373, 2388, 49, 21]
shared_interests_important: [1464, 2197, 1950, 1007, 126]
attractive: [131, 726, 131, 131, 726]
sincere: [57, 228, 57, 57, 228]
intelligence: [127, 409, 127, 127, 409]
funny: [19, 74, 1000, 19, 19]
ambition: [225, 697, 225, 225, 697]
attractive_partner: [284, 284, 284, 948, 948]
sincere_partner: [94, 94, 94, 353, 353]
intelligence_parter: [36, 36, 36, 193, 193]
funny_partner: [279, 279, 279, 733, 733]
ambition_partner: [119, 119, 119, 473, 473]
shared_interests_partner: [119, 119, 119, 473, 473]
sports: [650, 650, 650, 961, 961]
tvsports: [2151, 2151, 2151, 1292, 1292]
exercise: [619, 619, 619, 952, 952]
dining: [39, 39, 39, 172, 172]
museums: [117, 117, 117, 732, 732]
art: [224, 224, 224, 946, 946]
hiking: [963, 963, 963, 1386, 1386]
gaming: [2565, 2565, 2565, 2338, 2338]
clubbing: [912, 912, 912, 1068, 1068]
reading: [331, 331, 331, 331, 633]
tv: [1188, 1188, 1188, 1216, 1216]
theater: [288, 288, 288, 811, 811]
movies: [144, 462, 144, 144, 462]
concerts: [222, 222, 222, 777, 777]
music: [62, 62, 62, 196, 196]
shopping: [1093, 1093, 1093, 1098, 1098]
yoga: [2285, 2285, 2285, 1392, 1392]
interests_correlate: [2312, 985, 2312, 2597, 775]
expected_happy_with_sd_people: [321, 321, 321, 1262, 1262]
like: [273, 273, 273, 865, 865]

Output for 5_1.py:
Training accuracy: 0.78
Testing accuracy: 0.74

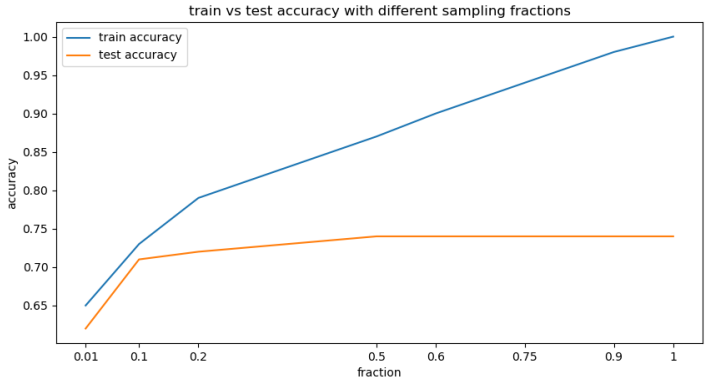
Output for 5_2.py:
bin size: 2
Training accuracy: 0.75
Testing accuracy: 0.71
bin size: 5
Training accuracy: 0.78
Testing accuracy: 0.74
bin size: 10
Training accuracy: 0.79
Testing accuracy: 0.76
bin size: 50
Training accuracy: 0.8
Testing accuracy: 0.76
bin size: 100
Training accuracy: 0.81
Testing accuracy: 0.76
bin size: 200
Training accuracy: 0.81
Testing accuracy: 0.75



The train accuracy will increase as bin size increases, but test accuracy stops increasing with too many bins. That's overfitting. So it's best to discrete continuous variables with bin size 10.

Output for 5_3.py:
fraction of sampling: 0.01
Training accuracy: 0.65
Testing accuracy: 0.62
fraction of sampling: 0.1

Training accuracy: 0.73
Testing accuracy: 0.71
fraction of sampling: 0.2
Training accuracy: 0.79
Testing accuracy: 0.72
fraction of sampling: 0.5
Training accuracy: 0.87
Testing accuracy: 0.74
fraction of sampling: 0.6
Training accuracy: 0.9
Testing accuracy: 0.74
fraction of sampling: 0.75
Training accuracy: 0.94
Testing accuracy: 0.74
fraction of sampling: 0.9
Training accuracy: 0.98
Testing accuracy: 0.74
fraction of sampling: 1
Training accuracy: 1.0
Testing accuracy: 0.74



The more data that we sample as training data, the better the accuracy is.