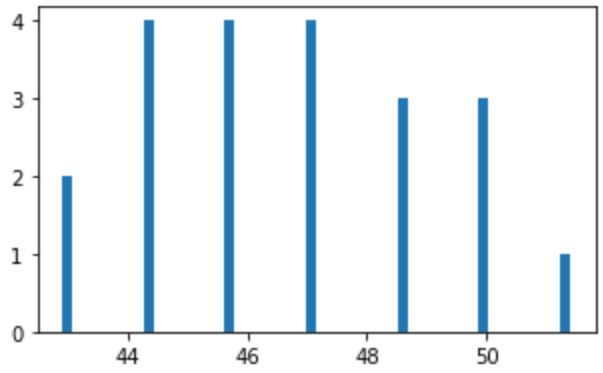
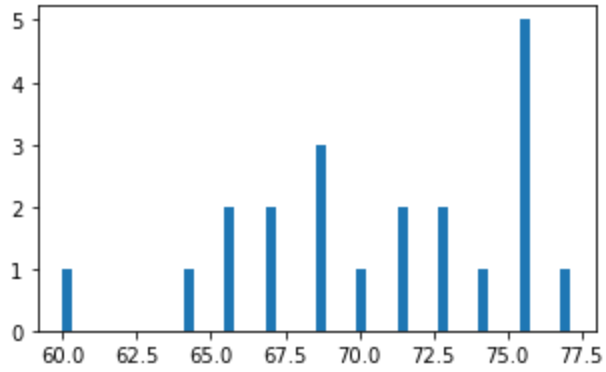
Three different custom heuristics were defined: (i) negative of the number of blank spaces on the board (defined below as custom_score_2), (ii) a random integer between -10 and 10 (defined below as custom_score_3), (iii) a linear combination of the previous two (this last heuristic actually worked the best, and was defined to be the first custom_score_1 and was the return value of the function custom_score()).

In order to evaluate the performance of the various heuristics, tournament.py was run a total of 21 times. The means and standard deviations of the win rates of the various heuristics were determined to be as follows:

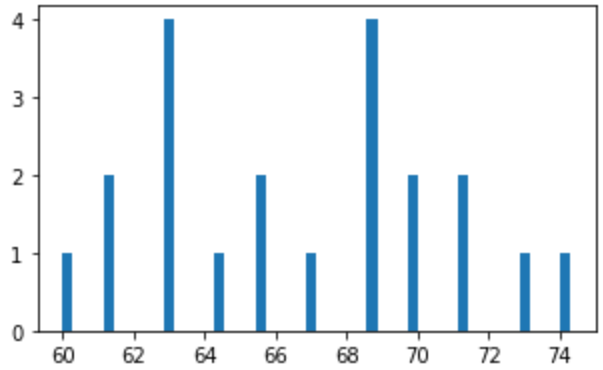
AB_Improved
Mean: 46.73
Std Dev: 2.41
Histogram of win rates below



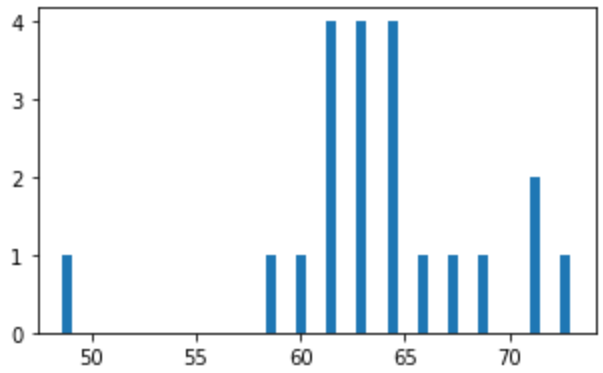
Custom_Score_1
Mean: 70.67
Std Dev: 4.54
Histogram of win rates below



Custom_Score_2
Mean: 66.74
Std Dev: 4.04
Histogram of win rates below



Custom_Score_3
Mean: 63.74
Std Dev: 5.05
Histogram of win rates below



Observing the means suggested that all three custom heuristics performed better than AB_Improved. In order to test this, a T-test was carried out between the mean win rates of each of the custom heuristics against that of AB_Improved. The p-values were found to be small (reported as follows), supporting the hypothesis that indeed all three performed better than AB_Improved.

Custom_Score_1: p-value ~ 10^-23
Custom_Score_2: p-value ~ 10^-21
Custom_Score_3: p-value ~ 10^-16

Of the three custom scores, custom_score_1 had the highest mean win rate. A similar T-test was carried out between the mean win rate of custom_score_1 against that of custom_score_2 and custom_score_3 to reveal that indeed, it performed better than the other two. The p-values for each of those two tests is reported below.

Custom_Score_2: p-value ~ 10^-3
Custom_Score_3: p-value ~ 10^-5

It is a bit of a surprise that a random score (custom_score_3) performed better than AB_Improved. However, using this observation along with the fact that the negative of total blank spaces left on the board also performed better than AB_Improved, the two scores were put together in a linear combination in the hope of producing a better performing score than both, which indeed it did.