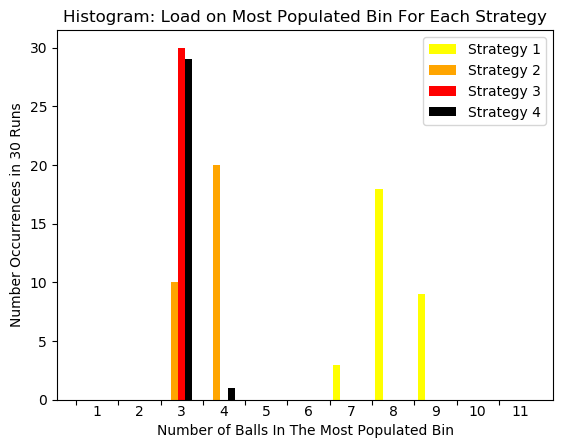
CS168 P1

Henry Lin

1. A. See code.py

B.



C.

For Strategy 1, it is the fastest to run because it just requires randomly choosing any bin. This does come with a cost though as you can observe the high relative collision rate in comparison to the other strategies. There is an improvement with Strategy 2, but both Strategy 3 and 4 seem to perform the best with few collisions on average compared to the others. Strategy 3 would be the slowest to run however, so Strategy 4 would be the sweet spot in terms of speed and number of collisions.

C. TODO ???? The larger X is, the longer the search time is since the entire linked list must be searched for an element.

D. TODO ??? Insertion time is always O(1) because you can put it at the beginning of the linked list. Search time will vary directly with X.

1. A. See code.py

B. Since there are 87925 total elements in the data stream, any heavy hitter must appear at least 880 times. This means there are 21 heavy hitters.

C. It does not affect the estimated counts. This is because no matter what the order of insertion, every relevant bucket in the table is incremented. This means that when the same elements are fed in, the end counts will turn out to be equal.

|  |  |  |
| --- | --- | --- |
|  | Number Heavy Hitters Average | Frequency of 9050 Average |
| Forward | 403.6 | 3215.6 |
| Reverse | 403.6 | 3215.6 |
| Random | 403.6 | 3215.6 |

D. See code.py

E. The count-min sketch will never underestimate the count of a value because every instance of a value will be hashed to the same bucket, meaning it will be at least equal to the actual frequency of the value in the data stream. The overestimation comes when collisions happen and different values are hashed to the same bucket, increasing the count to more than the actual.

F. It does affect the estimated counts. In this conservative implementation, not every relevant bucket is incremented and which bucket is incremented is contingent on the current count. This means that order of insertion will matter depending on which buckets fill up before others.

|  |  |  |
| --- | --- | --- |
|  | Number Heavy Hitters Average | Frequency of 9050 Average |
| Forward | 22.2 | 2864.9 |
| Reverse | 21.2 | 2500 |
| Random | 21.2 | 2500 |