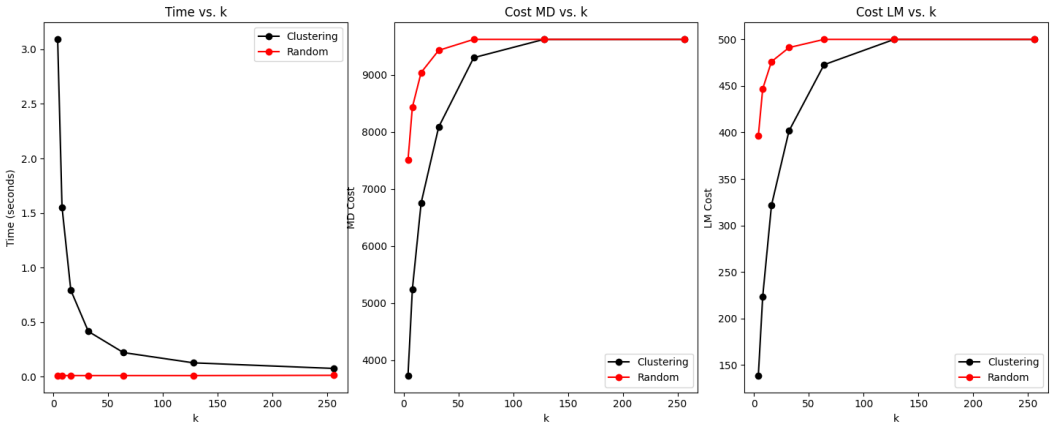
Comp430 Homework 1 Mini-Report, Batuhan Arat, 68665

I used mini-adult1 dataset to test my functions.




```
/usr/bin/python3 /Users/batuhanarat/Downloads/HW1/skeleton.py
k=256, Clustering Anonymizer-Time Cost: 0.07576918601989746 seconds, Cost MD: 9619, Cost LM: 500.0
k=256, Random Anonymizer-Time Cost: 0.012869834899902344 seconds, Cost MD: 9619, Cost LM: 500.0
k=128, Clustering Anonymizer-Time Cost: 0.12705111503601074 seconds, Cost MD: 9619, Cost LM: 500.0
k=128, Random Anonymizer-Time Cost: 0.01063394546508789 seconds, Cost MD: 9619, Cost LM: 500.0
k=64, Clustering Anonymizer-Time Cost: 0.22122502326965332 seconds, Cost MD: 9299, Cost LM: 472.8000000000013
k=64, Random Anonymizer-Time Cost: 0.0107362270355224461 seconds, Cost MD: 9619, Cost LM: 500.0
k=32, Clustering Anonymizer-Time Cost: 0.41529130935668945 seconds, Cost MD: 8083, Cost LM: 401.65221445221465
k=32, Random Anonymizer-Time Cost: 0.010702848434448242 seconds, Cost MD: 9427, Cost LM: 491.2000000000005
k=16, Clustering Anonymizer-Time Cost: 0.7928116321563721 seconds, Cost MD: 6755, Cost LM: 321.6632478632479
k=16, Random Anonymizer-Time Cost: 0.010875225067138672 seconds, Cost MD: 9039, Cost LM: 476.10000000000093
k=8, Clustering Anonymizer-Time Cost: 1.5529131889343262 seconds, Cost MD: 5243, Cost LM: 223.6139138639141
k=8, Random Anonymizer-Time Cost: 0.010722875595092773 seconds, Cost MD: 8431, Cost LM: 446.86616161616263
k=4, Clustering Anonymizer-Time Cost: 3.0907652378082275 seconds, Cost MD: 3731, Cost LM: 138.71587301587292
k=4, Random Anonymizer-Time Cost: 0.011126041412353516 seconds, Cost MD: 7503, Cost LM: 396.42426184926177
```

In the random anonymizer, when k is increasing, time cost is increased too. This is intuitive because it is harder to generate equivalence classes with higher members. While the k is getting large, cost of the generalization for each record is also getting larger. For md_cost and lmcost, while k is getting larger, generalizations are happening much more, so costs are increasing as well.

At the clustering anonymizer, when k is increasing, time cost is decreasing. At first this feels counterintuitive. I thought that if we are making higher generalization by increasing k, we should wait for more to program to execute, but actually vice versa is happening. This is happening because of the structure of the algorithm. When we are increasing the k, yes, we have more hypothetical cost to generalize k data in quasi-identifier, but we are also marking much record as used. The marked ones as used, does not taken again for the next record. So, we are decreasing the number of records to calculate hypothetical cost and that makes huge difference. In the graph it is seen that while k is increasing time will decreasing.

I don't complete the top-down algorithm. So, I only talk about random and clustering anonymizer. When I was working on the test cases, I saw that clustering algorithm is more efficient in terms of having less md and ld cost. But it is taking much more time. Random algorithm is gives higher md and ld costs, but it is faster. If I was working on the big dataset think I will choose clustering if I have time, unless I will choose random one. It is clear that random has a big time advantage, even that we cannot compare in the time vs k graph because it is too small.