In this analysis, I tried to analyze different types of anonymization techniques. I conduct my analysis with different k and seed values. All the graphs are created with several runs. For more details https://github.com/EsadSimitcioglu/COMP-530

# Random Anonymizer

In this technique, we try to create Equivalence Classes with neighbor rows. First, we shuffle the whole dataset after that we grouped the rows by k number. This technique is a more basic approach than the other techniques that I used. So, the implementation of this anonymizer is lot easier than the others. If we compare it according to the working time, we see that it works faster than the others. Depending on complexity and runtime, this technique outperforms other techniques in complexity and runtime. This technique gives a poor performance in increasing the K value when we talk about the distortion and loss metric. As we increase K, the time complexity decreases. But this reduction is not that big. I created the chart with more precise values for greater precision. If I had less stature in the Domain Generalization Hierarchy, I would use random anonymizer. That way, it doesn't need to do any more calculations as the values aren't too different. Even if we must generalize the QI values, it won't take much time.

# Clustering Anonymizer

In this technique, we try to create Equivalence Classes with less lost metric cost. First, we select the top records from the dataset, then from the dataset we find the k-1 record that has less lost metric cost than our first selected record. The implementation was a little more difficult than the previous approach. However, the results seem more satisfactory. Due to complexity and uptime, this technique lags a bit over the previous one. For all K values, Clustering costs less for both metrics. This is because of the metric analysis we did when creating the Equivalence Classes. The distortion and loss metric costs are seen least in this technique. If I have more diverse values in the dataset or If I want to achieve lower k-anonymity, Clustering Anonymizer will be a better choice. Because building EC with minimal lost metric cost would be a better choice for utility. Clustering takes more time than random anonymizer, but costs less lost metrics for less k-value.

## Conclusion

In this homework, I kind an experienced what I think that I would. Randomness is sometimes good for anonymization but it’s not the best option. It’s far easier to implement but the utility is very poor. I experienced generalization techniques for the first time. The implementation of these techniques is not easy as we seen in the class. When I look back it looks easier to implement these techniques. That’s why I’m happy to do homework like this. It was one of the most enjoyable homework assignments for me.