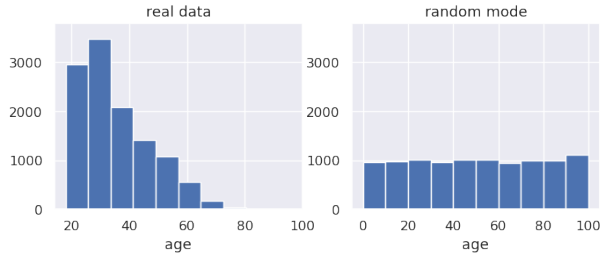
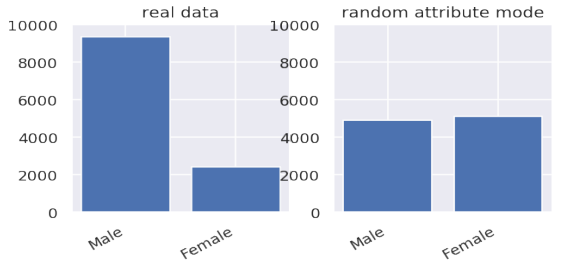
Goal 1)
Q1.

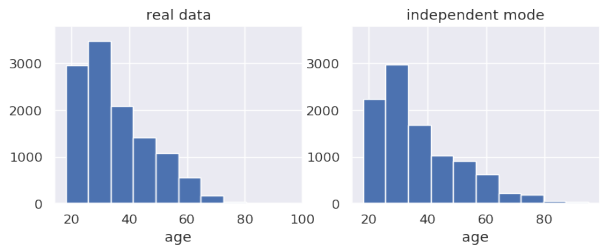
Hw_compass	Median	Mean	Min	Max
A_Age	50.0	50.3008	0	100
B_Age	33.0	37.146	18	95
C_Age	32.0	35.4044	18	95
D_Age	32.0	35.1127	18	87
A_Score	5.0	4.9073	-1	10
B_Score	4.0	4.3271	-1	10
C_Score	4.0	4.3769	-1	10
D_Score	4.0	4.4378	-1	10
Groud Truth Age	32	35.1433	18	96
Groud Truth Score	4.0	4.3712	-1	10

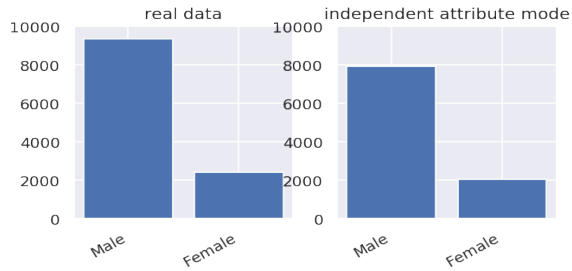
We can see that mode ***A***, the ***random mode***, has the worst accuracy as compared to the ground truth, especially for the 'age' feature. All other modes perform relatively well when compared to the ground truth for both features. The performance of the random mode is worse off because the data is sampled randomly, and thus it does not conform very well to the ground truth.

Q2.







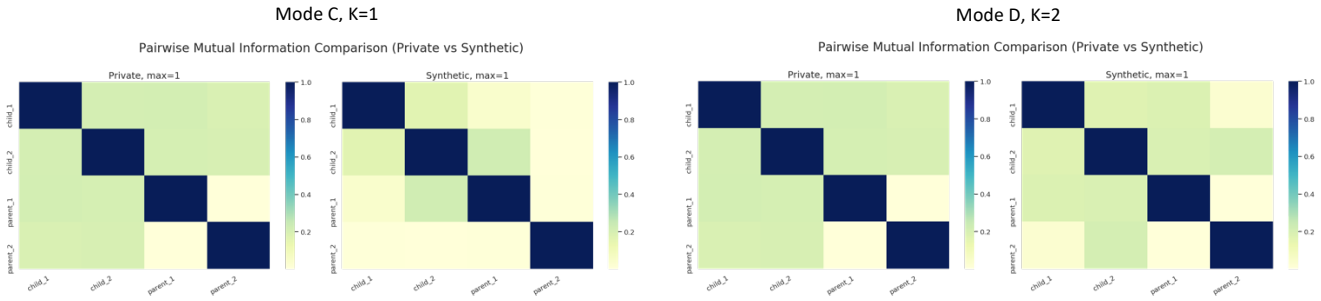


From the above plots, we can see that the ***independent mode***, ***B***, is more similar to the real data than the ***random mode***, ***A***. This is because the random mode is uniformly sampling from the real data whereas the independent mode is sampling from the attributes histogram, guaranteeing the similarity to the real data.

Probability Distribution	KS test (age)	KL divergence (sex)
Random (A)	0.3734	0.2239
Independent (B)	0.0518	1.2082

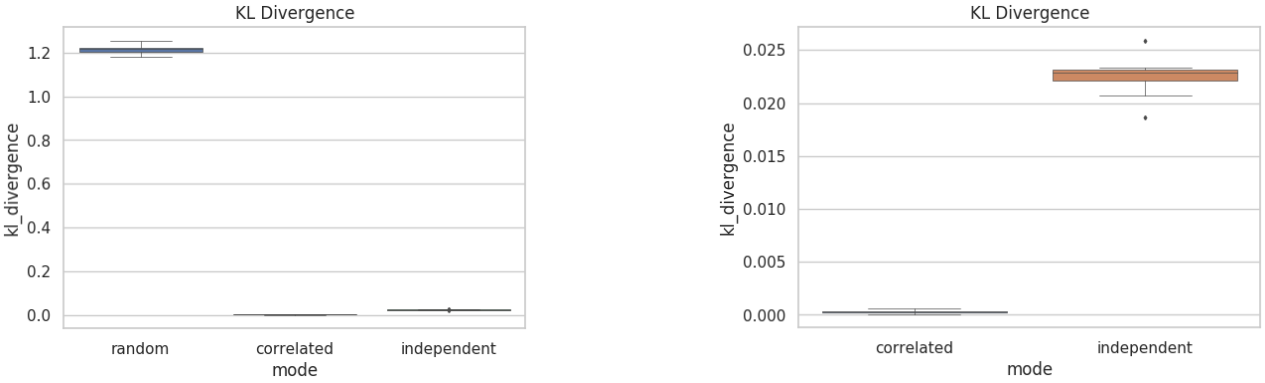
Mode ***B, independent mode***, has lower KS test score and higher KL divergence than mode ***A, random mode***. This is because mode B is more similar to the ground truth, resulting in lower KS test (which measures how much the two underlying probability distribution differs), and higher KL divergence. The contrary applies to mode A for its difference from the real data.

Q3.



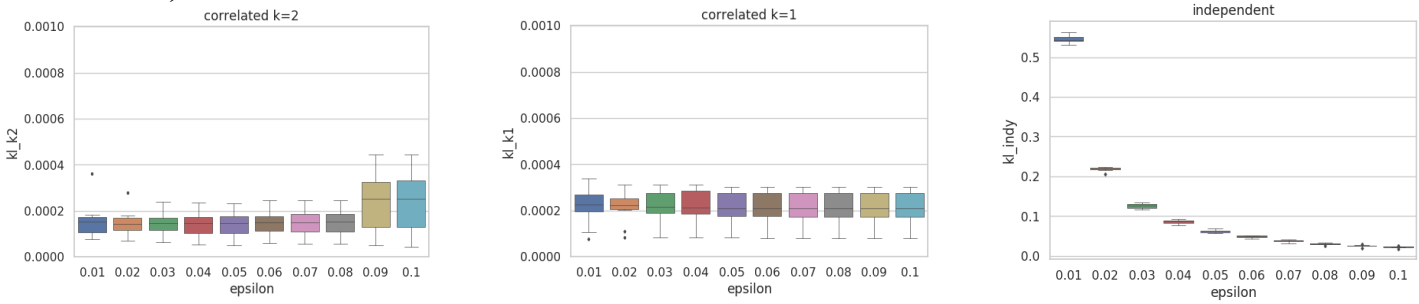
We can see for ***mode D*** where K=2 has a more similar heat map to the private/true data set. This means that mutual information is better preserved in mode D, except for the 'parent_2 & child_2' mutual information.

Goal 2)



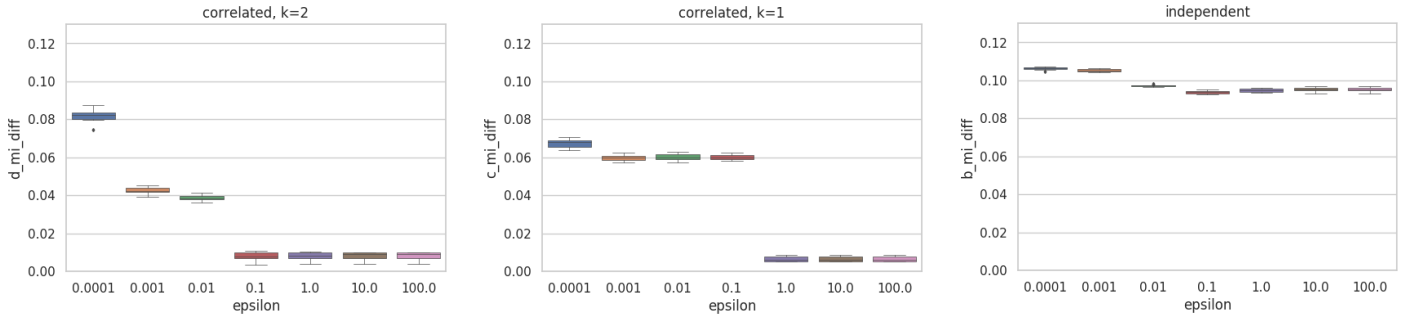
Distribution-wise, we can see the ***random mode (A)*** has wider distribution (higher variability) than ***independent (B)*** mode which in turn has wider distribution than ***correlated (C)*** mode. On the other hand, mode A sits at around 1.2 for KL divergence when mode B (between 0.02 and 0.025) and C (close to 0) have lower values. Higher KL divergence indicates more difference in probability distribution. This suggests that the probability distribution for mode A is more different from the original/private dataset.

Goal 3)



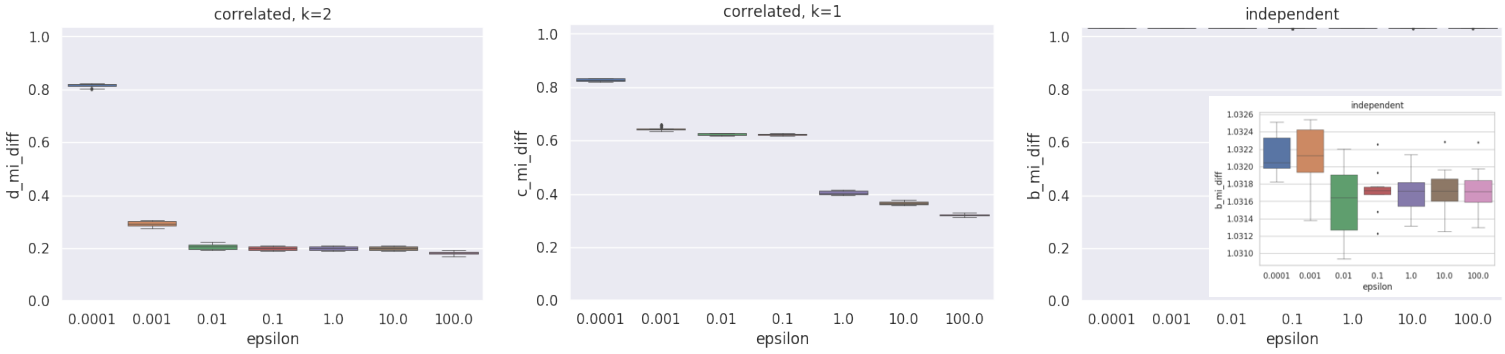
From the plots above, we can see that, for the **Race attribute**, the correlated mode C and D have relatively stable performance over a range of epsilon. For the independent mode B, however, lower epsilon results in higher KL divergence. This is because lower epsilon leads to better privacy protection and thus a higher degree of difference in the synthetic probability distribution. As epsilon grows larger, KL divergence gets closer to 0. It is also worth notice that independent mode has higher KL divergence at the same level of epsilon compared to mode C and D, suggesting that the independent generator creates more different probability distribution than the original data.

Compas Data



For the ***Compas Data***, we can see generally higher epsilon leads to lower difference in mutual information, with independent mode (B) having higher difference overall at the same epsilon level. A higher difference in mutual information means that the synthetic data is more different than the private/original dataset. This makes sense because lower epsilon values lead to better privacy preservation, and thus more difference between the datasets.

Fake Data



For the ***Fake Data***, we can see generally higher epsilon leads to lower difference in mutual information, with independent mode (B) having higher difference overall at the same epsilon level. From the zoom-in view on mode B, we can see that although it has higher mutual information difference overall, the general inverse trend between epsilon and mutual information difference still holds. A higher difference in mutual information means that the synthetic data is more different than the private/original dataset. This makes sense because lower epsilon values lead to better privacy preservation, and thus more difference between the datasets.