DS-UA 202: Homework 3

Daphne Ozkan

April 27, 2025

Problem 1

Part (a)

Differential privacy requires that the probability of a response does not depend significantly on an individual's data, ensuring that the privacy of individuals is preserved. The randomized response mechanism described here does not satisfy differential privacy because it does not guarantee plausible deniability for all inputs. Specifically, if we observe the answer as "yes", we cannot be certain whether the truth is "yes" or "no". However, if the response is "no", we can be certain that the true answer is "no" because the probability of the answer being "no" given that the true answer is "yes" is zero.

We can further illustrate this by calculating the conditional probabilities for each scenario:

$$P(\text{Response} = \text{Yes} \mid \text{Truth} = \text{Yes}) = 1, \quad P(\text{Response} = \text{Yes} \mid \text{Truth} = \text{No}) = 0.5$$

From this, the epsilon (ϵ) can be calculated using the definition of differential privacy:

$$\epsilon = \ln\left(\frac{P(\text{Response} = \text{Yes} \mid \text{Truth} = \text{Yes})}{P(\text{Response} = \text{Yes} \mid \text{Truth} = \text{No})}\right) = \ln\left(\frac{1}{0.5}\right) = \ln(2) = 0.693$$

However, if the response is "no", the differential privacy parameter becomes problematic:

$$P(\text{Response} = \text{No} \mid \text{Truth} = \text{Yes}) = 0, \quad P(\text{Response} = \text{No} \mid \text{Truth} = \text{No}) = 0.5$$

$$\epsilon = \ln\left(\frac{P(\text{Response} = \text{No} \mid \text{Truth} = \text{No})}{P(\text{Response} = \text{No} \mid \text{Truth} = \text{Yes})}\right) = \ln\left(\frac{0.5}{0}\right) = \ln(\infty)$$

In this case, ϵ becomes infinite, which means that the mechanism does not preserve privacy when the answer is "no". This results in the complete loss of privacy, as the response can fully reveal the truth.

Thus, this mechanism does not satisfy the differential privacy definition, as it allows us to completely deduce the individual's true answer in the case where the response is "no". Therefore, the randomized response mechanism in this scenario is not differentially private, and the calculation of epsilon does not hold for all possible answers.

Part (b)

Truth	**P(Response = Yes | Truth)**	**P(Response = No | Truth)**
Yes	$\frac{1}{2} + \frac{1}{2} \times \frac{1}{3} = \frac{2}{3}$	$\frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$
No	$0 + \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$	$\frac{1}{2} + \frac{1}{2} \times \frac{2}{3} = \frac{5}{6}$

Now we can calculate the value of epsilon for each scenario:

First, we calculate the epsilon value for "Truth = Yes":

$$\epsilon = \ln\left(\frac{P(\text{Response} = \text{Yes}|\text{Truth} = \text{Yes})}{P(\text{Response} = \text{Yes}|\text{Truth} = \text{No})}\right) \implies \epsilon = \ln\left(\frac{\frac{2}{3}}{\frac{1}{6}}\right) = \ln(4) \approx 1.386$$

Next, Calculate the epsilon for "Truth = No":

$$\epsilon = \ln\left(\frac{P(\text{Response} = \text{No}|\text{Truth} = \text{No})}{P(\text{Response} = \text{No}|\text{Truth} = \text{Yes})}\right) \implies \epsilon = \ln\left(\frac{\frac{5}{6}}{\frac{1}{3}}\right) = \ln\left(\frac{5}{2}\right) \approx 0.916$$

Now that we have calculated epsilon for both scenarios:

- $\epsilon_{\text{Yes}} = \ln(4) \approx 1.386$
- $\epsilon_{\text{No}} \ln(\frac{5}{2}) \approx 0.916$

The worst-case scenario is the scenario with the highest epsilon, which in this case is when the truth is "Yes" with $\epsilon = \ln(4) \approx 1.386$. Therefore, the differential privacy parameter for this mechanism is $\epsilon = \ln(4) \approx 1.386$, representing the worst-case scenario for privacy, which indicates that the mechanism does not provide optimal privacy protection in this case.

Problem 2

Part (a)

Question 1

The table below presents the median, mean, min, and max values for variables age and score in the real dataset (hw_compas) and synthetic datasets generated under modes A, B, C, and D.

	hw_compas		A		B		C		D	
	age	score	age	score	age	score	age	score	age	score
median	32.000	4.000	51.000	5.000	33.000	4.000	36.000	5.000	39.000	4.000
mean	35.143	4.371	50.173	4.939	35.735	4.366	41.579	4.949	44.153	4.466
min	18.000	-1.000	0.000	-1.000	18.000	1.000	18.000	-1.000	18.000	-1.000
max	96.000	10.000	100.000	10.000	76.000	10.000	96.000	10.000	96.000	10.000

The table below presents the differences in the statistics (median, mean, min, max) between the real data (hw_compas) and synthetic datasets under modes A, B, C, and D. The first two columns under hw_compas are the statistics for the real dataset, and the rest of the columns under modes A, B, C, and D represent the difference in values between the synthetic dataset and the real dataset, so we can directly see which synthetic datasets generated values that deviated the most from the real dataset.

	hw_compas		A		B		C		D	
	age	score	age	score	age	score	age	score	age	score
median	32.000	4.000	-19.000	-1.000	-1.000	0.000	-4.000	-1.000	-7.000	0.000
mean	35.143	4.371	-15.030	-0.568	-0.592	0.006	-6.435	-0.577	-9.010	-0.095
min	18.000	-1.000	18.000	0.000	0.000	-2.000	0.000	0.000	0.000	0.000
max	96.000	10.000	-4.000	0.000	20.000	0.000	0.000	0.000	0.000	0.000

In the above two tables, we can see that the difference in the median age between the real dataset and the synthetic datasets is quite large for some modes. For instance, in Mode A (random mode), the median age is 51, which is a large deviation (-19.000, as we can see in the second table), from the median age in the real dataset, which is 32. Modes B, C, and D show less extreme deviations, with Mode B (independent attribute mode) having the closest median age to the real dataset. The median age under Mode B is 33, which is a deviation of -1.000 in the second table). The median score of the real dataset is 4, which is close to the median scores generated by the synthetic datasets: Mode A has

a median score of 5, Mode B has an median score of 4, Mode C has a median score of 5, and Mode D has a median score of 4. Therefore Modes B and D have the same median score as the real dataset, and Modes A and C only have a deviation of -1.000.

Next, we also see some deviation in the values for mean age and score between the real dataset and synthetic datasets. The mean age in the dataset is approximately 35.143, which is close to the mean age in Mode B (which is 35.735). The mean ages for the rest of the modes are further off, with Mode A having the largest deviation once again since it has a mean age of 50.173, followed by Mode D, which has a mean age of 44.153, and then Mode C, which has a mean age of 41.579. The mean score of the real dataset is 4.371, which is closest to the mean score of Mode B (which has a mean score of 4.366), followed by the mean score of Mode D (4.466), then Mode A (4.939) and lastly Mode C (4.949).

The minimum age for the real dataset is 18, which is the same as the minimum ages for Modes B, C, and D. Mode A has a minimum age of 0.000, so it is the furthest off from the minimum age in the real dataset. The minimum score for the real dataset is -1.000, which is the same as the minimum score for Modes A, C, and D. Mode B has a minimum score of 1.000, which is not that far off from the minimum score of the real dataset.

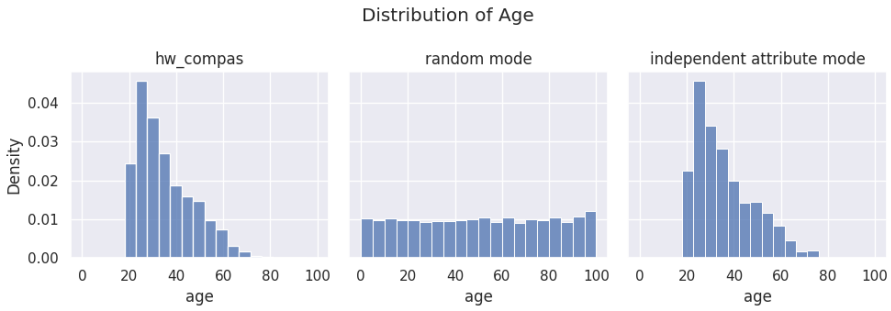
The maximum age for the real dataset is 96, which is the same as the maximum age for Modes C and D. Mode A has a maximum age of 100, and Mode B has a maximum age of 76, which is the farthest off from the maximum age of the real dataset (a deviation of 20.000). The maximum score for the real dataset is 10, which is the same as the maximum score for Modes A, B, C, and D, so all of the synthetic datasets were able to accurately reflect the maximum score of the real dataset.

Based on these statistical comparisons, we can see that Mode B (Independent Attribute Mode) is overall the most accurate in replicating the real dataset's statistical properties. When comparing the median, mean, min, and max values for both age and score attributes, we found that Mode B is generally the closest to the values in real dataset (with the exception of minimum score which had a very small deviation of -2.000, and maximum age, which had a deviation of 20.000). Mode A (Random Mode) is the least accurate, with the largest overall deviations from the real dataset. Mode C (Correlated Attribute Mode, k=1) and Mode D (Correlated Attribute Mode, k=2) show moderate accuracy, as both modes show improved results compared to Mode A, but still exhibit noticeable differences from the real data.

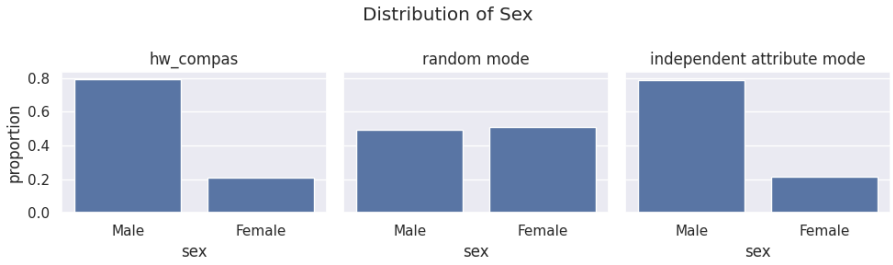
The substantial differences in accuracy arise from the nature of each mode's data generation method. Mode B (Independent Attribute Mode) is the most accurate method because it assumes independence between attributes, which al-

lows it to generate synthetic data that closely matches the real data's statistical properties. Mode A (Random Mode) is the least accurate because it generates synthetic data randomly, without any regard for the relationships between the attributes. This leads to large deviations from the statistical values in the real dataset, especially for the age attribute. Modes C (Correlated Attribute Mode, k=1) and D (Correlated Attribute Mode, k=2) fall in the middle. While these modes introduce correlations between attributes, they still have noticeable deviations from the real dataset. Mode D, with a higher correlation degree (k=2), preserves the correlations between attributes more strongly than Mode C, but both modes exhibit similar levels of accuracy. They better preserve relationships between attributes compared to Mode A, but still show more deviation from the real data than Mode B, especially in terms of the median and mean values for age and score.

Question 2



In the above set of histograms, we can see the distribution of the age attribute in both the real dataset and the synthetic datasets under random mode (A) and independent attribute mode (B). The real dataset has a distribution that heavily peaks around younger ages, specifically in the range of around 18-30. In random mode (A), on the other hand, the distribution of age is more uniformly spread across the entire age range, with no clear peak, indicating a substantial loss in replicating the true distribution of the real dataset. This shows that when attributes are generated randomly, they fail to preserve the natural age distribution in the data, leading to a poor representation. In contrast, the independent attribute mode (B) does a better job of preserving the true distribution. The distribution of age in Mode B is very close to the distribution of the real dataset, with a similar shape and a clear peak around the lower age range. This suggests that independent attribute mode is more accurate in replicating the real data's statistical properties, compared to random mode.



For the above set of bar charts that show the distribution of the sex attribute, the bar charts show the proportion of male and female values in each dataset. In the real dataset, the majority of the data consists of males (approximately 80%), with a small proportion of females (around 20%). In random mode (A), the proportion of male to female is much less skewed and nearly equal, leading the distribution of sex to become more balanced, suggesting that the random generation of the sex attribute fails to preserve the real distribution of the attribute. Independent attribute mode (B) performs better, where the proportion of male to female matches the original dataset much more closely than random mode (A).

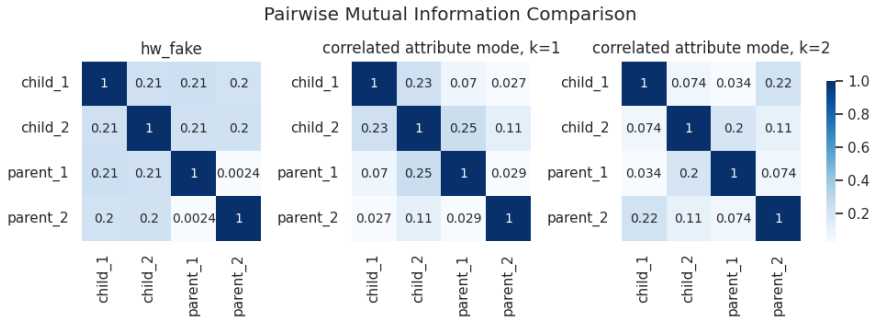
	KS test (age)	KL-Divergence (sex)
random mode	0.373509	0.223198
independent attribute mode	0.026252	0.000249

In the above table, we can see that the KL-Divergence for the sex attribute is significantly lower in independent attribute mode B (0.000249) compared to random mode A, which is 0.223198, indicating that Mode B does a much better job of preserving the original distribution of the sex attribute. The KS test results for age show that independent attribute mode B has a much lower score (0.026252) than Mode A (0.373509), meaning that Mode B is statistically closer to the real data than Mode A, confirming its better performance in preserving the distribution of both age and sex.

In conclusion, Mode B (independent attribute mode) is the most accurate in replicating the original distributions of age and sex, as it preserves the statistical properties of the real dataset more effectively than Mode A (random mode). Mode A generates synthetic data randomly without considering any correlations between attributes, leading to significant deviations from the real dataset. Mode B assumes independence between attributes, which allows it to better capture the distributions of both age and sex.

Question 3

The heatmap below visualizes the pairwise mutual information between the attributes in the hw_fake dataset, as well as in the synthetic datasets generated under Mode C (k=1) and Mode D (k=2).



When comparing the pairwise mutual information values in the original dataset (hw_fake) to Mode C (correlated attribute mode, k=1) and Mode D (correlated attribute mode, k=2), I noticed the following differences:

In Mode C (correlated attribute mode, k=1), the mutual information between certain attributes is similar to the values in the original dataset, hw_fake. For example, the mutual information between child_1 and child_2 is 0.23 in Mode C, which is similar to the mutual information between child_1 and child_2 in hw_fake (0.21). Additionally, the mutual information between attributes child_2 and parent_1 is 0.25 in Mode C, which is close to the mutual information value of 0.21 between child_2 and parent_1 in hw_fake. However, the mutual information values between other attributes in Mode C are not as close to the mutual information values in the original dataset, hw_fake. For example, the mutual information values between attributes parent_1 and child_1 is 0.07 in Mode C, while the mutual information values for parent_1 and child_1 is 0.21 in hw_fake. Furthermore, the mutual information value for parent_2 and child_1 0.027 in Mode C, while the mutual information values for parent_2 and child_1 is 0.2 in hw_fake. Also, the mutual information values between parent_2 and child_2 is 0.11 in Mode C, which is a moderate deviation from the mutual information value between parent_2 and child_2 in hw_fake (which is 0.2). Lastly, the mutual information between the attributes parent_2 and parent_1 is 0.029 in Mode C, whereas the mutual information value between parent_2 and parent_1 is 0.0024 in hw_fake, which is a larger deviation from the original dataset.

Similarly, in Mode D (correlated attribute mode, k=2), the mutual information between some attributes is similar to the values in hw_fake, whereas the mutual information between other attributes is not preserved as well in the synthetic dataset. For example, the mutual information between attributes child_2 and parent_1 is 0.2 in Mode D, which is nearly identical to the mutual information value of 0.21 between child_2 and parent_1 in hw_fake. Additionally, the mutual information values between child_1 and parent_2 is 0.22 in Mode D, which is very close to the mutual information values between child_1 and parent_2 in hw_fake (which is 0.2). However, the mutual information values between other attributes in Mode D are not as close to the mutual information values in the original dataset, hw_fake. For example, the mutual information between child_1 and child_2 is 0.074 in Mode D, which is a somewhat large deviation from the mutual information between child_1 and child_2 in hw_fake (0.21). Furthermore, the mutual information value for parent_1 and child_1 0.034 in Mode D, while the mutual information values for parent_1 and child_1 is 0.21 in hw_fake. Also, the mutual information values between parent_2 and child_2 is 0.11 in Mode D, which is a moderate deviation from the mutual information value between parent_2 and child_2 in hw_fake (which is 0.2). Lastly, the mutual information between the attributes parent_2 and parent_1 is 0.074 in Mode C, whereas the mutual information value between parent_2 and parent_1 is 0.0024 in hw_fake, which is a larger deviation from the original dataset.

In conclusion, both Mode C and Mode D succeed in preserving the relationships between certain attributes, but neither perfectly replicates the mutual information values seen in the original dataset. Mode D, with its higher correlation degree (k=2), is generally better at preserving some dependencies, although Mode C performs better in some cases, such as with the child_1 and child_2 relationship.

	Mutual Information Distance
correlated attribute mode, k=1	0.131363
correlated attribute mode, k=2	0.169164

The table above displays the Mutual Information Distance for the synthetic datasets generated under Modes C (k=1) and D (k=2), which quantifies the difference in mutual information between the real dataset (hw_fake) and the synthetic datasets. A smaller distance indicates that the synthetic data preserves the mutual information better.

From the table, we can see that Mode C (k=1) has a Mutual Information Distance of 0.131363, which is relatively moderate. This indicates that Mode C does a fairly good job of preserving the mutual information between the real data and the synthetic data, but there are still noticeable discrepancies, especially for certain attribute pairs where the mutual information is not well-preserved.

Mode D (k=2) has a Mutual Information Distance of 0.169164, which is slightly higher than Mode C. This suggests that Mode D, despite its higher correlation degree, introduces more deviations when compared to the real dataset. The stronger correlations in Mode D may help preserve certain relationships better, but it also introduces more complexity, leading to slightly worse overall preservation of mutual information.

Part (b)

The table below represents the values for the median, mean, min, and max of age for the real dataset (hw_compas) and 10 synthetic datasets generated under modes A, B, and C, with a fixed epsilon = 0.1 and 10 different seeds, ranging from 0 to 9. The synthetic datasets were generated by specifying different seeds, which allows for the variability in the results observed across these datasets.

	hw_compas	A	B	C
median	32.000000	50.200000	32.200000	35.900000
mean	35.143319	50.10166	36.11932	40.94772
min	18.000000	0.000000	18.000000	18.000000
max	96.000000	100.000000	88.500000	94.800000

The second table below presents the differences between the values for median, mean, min, and max of age in the synthetic datasets generated under Modes A, B, and C, compared to the corresponding values in the real dataset (hw_compas). These differences are calculated by subtracting the real dataset values from the synthetic dataset values, as shown in the first table above. This table highlights the deviations between the synthetic data and the real data for each of the modes, allowing for a better understanding of how accurately each mode replicates the statistical properties of the real dataset.

	hw_compas	A	B	C
median	32.000000	-18.200000	-0.200000	-3.900000
mean	35.143319	-14.958341	-0.976001	-5.804401
min	18.000000	18.000000	0.000000	0.000000
max	96.000000	-4.000000	7.500000	1.200000

Based on the two tables above, I noticed the following:

Mode A (Random Mode) shows the largest deviations from the real data. The median of Mode A in the first table is 50.2, which is a large deviation from the real dataset's median of 32. Similarly, the mean of Mode A, which is approximately 50.10, is far from the real dataset's mean of approximately 35.143. The minimum value of Mode A is 0, which is also much lower than the real dataset's minimum of 18, while the maximum value of Mode A is 100, which is relatively higher than the real dataset's maximum of 96. This suggests that Mode A, which generates synthetic data randomly without accounting for the relationships between attributes, produces data that deviates significantly from the real dataset.

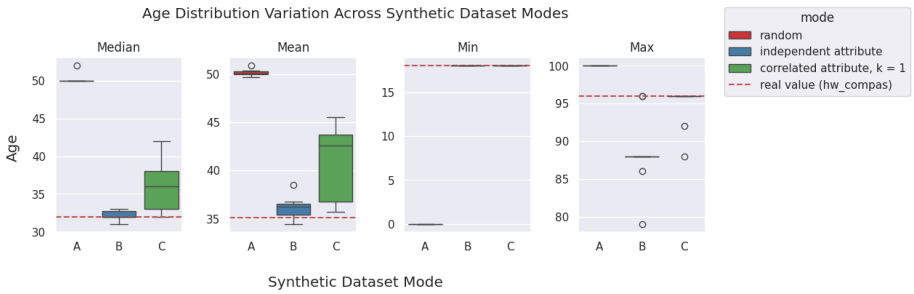
Mode B (Independent Attribute Mode) produces the most accurate synthetic datasets overall. The values for Mode B's median and mean age are very close to the median and mean age of the real dataset, and the value for minimum age in Mode B is identical to that of the real dataset. The only notable deviation is the value for maximum age in Mode B, which is 88.5, compared to a maximum age of 96 in the real dataset. This indicates that Mode B is successful in maintaining the statistical distribution of the real data while generating synthetic data.

Mode C (Correlated Attribute Mode, k=1) provides moderately accurate results, but it deviates more from the real data compared to Mode B. The median

age of Mode C is 35.9, which is still reasonably close to the real dataset's median (32), but it shows a slightly larger deviation. The mean is also higher in Mode C (40.95), indicating that Mode C generates synthetic data with a relatively higher average age. The minimum age in Mode C is identical to the minimum age of the real dataset, and the maximum age in Mode C is 94.8, which is close to the real dataset's maximum age of 96. Although Mode C does introduce some correlations between attributes, it does not preserve the data's statistical properties as well as Mode B.

Thus, Mode A exhibits the most variability across the 10 synthetic datasets, with significant fluctuations median, mean, and minimum age, as well as a moderate deviation in maximum age. Mode B shows the least variability, as its synthetic datasets remain relatively close to the real dataset's statistical values. Mode C shows moderate variability, with some fluctuations in the mean and median but less than Mode A.

In conclusion, Mode B is the most accurate in preserving the statistical properties of the real dataset, with Mode C following closely behind. Mode A, due to its random data generation method, shows the largest deviations from the real data and exhibits the most variability. The results highlight that using independent attributes (Mode B) preserves the statistical properties of the data more effectively than using random or correlated attribute methods.



The box plots above provide a visual representation of the distribution of the synthetic datasets generated under Modes A, B, and C, with the real dataset (hw_compas) as a reference represented by the red dotted line in each plot. The box plot allows for a clear comparison of the variability and accuracy of the synthetic datasets across the different modes for the median, mean, min, and max values of age.

From these box plots, I noticed the following:

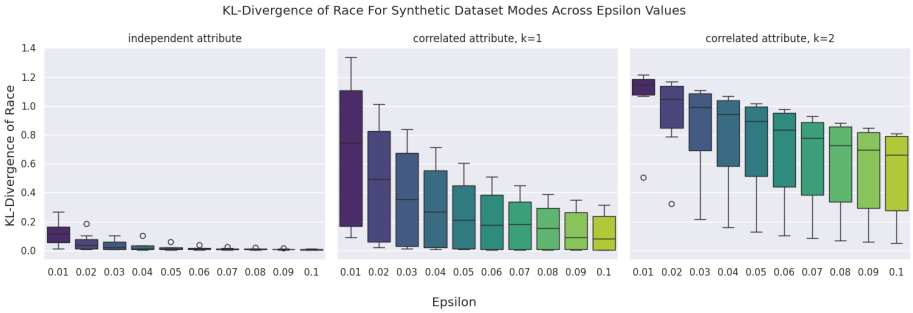
The distribution of Mode A is farthest from the red dotted line, which represents the distribution of hw_compas, acrosss median, mean, and minimum age values. It is also relatively far from the red dotted line in the maximum age boxplot. This therefore confirms my observations from the tables above, which is that Mode A exhibits the greatest variability among synthetic modes A, B, and C. This is because Mode A has a wide spread of values for the median, mean, min, and max which reflects the random nature of the synthetic data generation.

The distribution of Mode B (Independent Attribute Mode) is closest to the red dotted line, which represents the distribution of hw_compas, in median, mean, and minimum age values. For maximum age, Mode B is furthest from the red dotted line, indicating that the distribution for maximum age in Mode B is not as closely aligned with the distribution of maximum age in the real dataset, although this is only a moderate deviation from the real dataset. It therefore has the least variability among the three synthetic dataset modes shown, as its statistical values overall closely align with the real dataset. The plot also shows a narrow range for the synthetic data in Mode B, indicating that Mode B generates synthetic datasets that consistently replicate the statistical properties of the real data.

The distribution of Mode C (Correlated Attribute Mode, k=1) has a broader range for median and mean age values, as shown in the first two box plots, but is more closely aligned with the red dotted line for the minimum and maximum age values. Based on this, we can see that Mode C therefore falls in between Modes A and B in terms of variability. While it shows less variability than Mode A, the range of values is still wider than Mode B, indicating that Mode C does not preserve the real dataset's statistical properties as accurately as Mode B. Mode C's distribution is more concentrated around the real dataset's values but still deviates more than Mode B.

In conclusion, the box plot confirms that Mode B is the most accurate with the least variability, Mode A is the least accurate with the most variability, and Mode C provides intermediate results with more variability than Mode B but less than Mode A.

Part (c)



The box-and-whiskers plots above visualize the KL-divergence of the race attribute among the synthetic datasets generated under Modes B, C, and D at different epsilon values (ranging from 0.01 to 0.1). These plots allow us to observe how well the synthetic datasets preserve the distribution of the race attribute as the privacy budget (epsilon) changes.

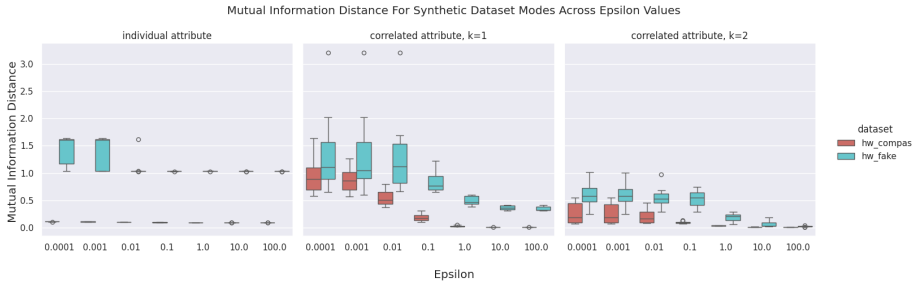
Based on these plots, my findings are as follows:

Mode B (Independent Attribute Mode): The plot for Mode B shows the lowest KL-divergence scores overall, particularly for lower epsilon values because the attributes are generated independently, making it easier for the synthetic data to match the real dataset's distribution. However, this method does not account for any correlations between attributes, leading to the best privacy-preserving results with minimal divergence from the original dataset, indicating that this method achieves a better balance between privacy and maintaining the integrity of the original data.

Mode C (Correlated Attribute Mode, k=1): The plot for Mode C shows a moderate KL-divergence compared to Mode B, with KL divergence decreasing as epsilon increases, indicating that the synthetic datasets become more similar to the original data's race distribution with higher privacy budgets. However, compared to Mode B, the KL-Divergence values in Mode C are higher across all epsilon values, suggesting that Mode C does not preserve the race distribution of the original dataset as effectively as Mode B. This is because while correlations between attributes are introduced, the strength of the correlation is relatively low (k=1), which leads to a less accurate replication of the real dataset's distribution of race compared to Mode B.

Mode D (Correlated Attribute Mode, k=2): The plot for Mode D shows a higher KL-divergence compared to Mode B and C. This trend is consistent across all epsilon values, with Mode D displaying higher variability in the divergence. As epsilon increases, the KL-Divergence decreases, indicating that the privacy-preserving features are having a greater effect in reducing the divergence as the privacy budget increases. However, even at higher epsilon values, the range of KL-Divergence values grows (i.e., the box plots towards the left at smaller epsilon values have smaller, shorter ranges and the box plots towards the right have larger, longer ranges of KL divergence values). This increase in variability can be explained by the relationship between epsilon and the privacy budget. At smaller epsilon values (such as 0.01), differential privacy introduces higher levels of noise to the synthetic data in order to ensure stronger privacy protection. The result is a more constrained dataset with less variation, reflected in a narrower boxplot and lower variability in the KL divergence values. On the other hand, as epsilon increases, the noise added to the synthetic data decreases, allowing the synthetic datasets to resemble the real data more closely, but also introducing more variability. This weaker privacy protection (at higher epsilon values) leads to greater variability in the synthetic datasets, as evidenced by the longer bar plots as the epsilon values increase up to 0.1.

Overall, Mode B achieves the lowest KL-divergence due to its independent attribute generation, balancing privacy and data accuracy. Mode C introduces correlations but with a weaker preservation of the race distribution, while Mode D, despite stronger correlations, shows the highest variability, especially as epsilon increases, indicating a trade-off between privacy and data fidelity.



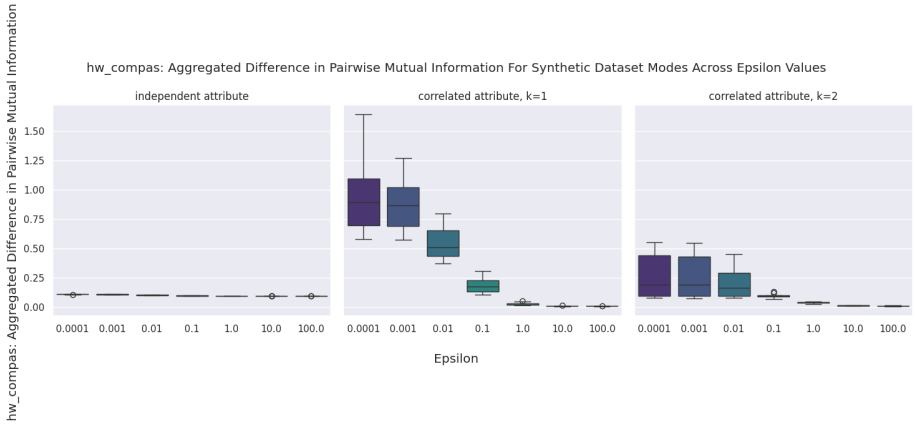
The box-and-whiskers plots above show the Mutual Information Distance for synthetic datasets generated under different modes (individual attribute, correlated attribute with k=1, and correlated attribute with k=2) across varying epsilon values. These plots measure how closely the synthetic datasets replicate the mutual information properties between pairs of attributes in comparison to the real data (hw_compas), with lower distances indicating better preservation of inter-attribute relationships.

Mode B (Independent Attribute Mode): Compared to the other plots representing Modes C and D, Mode B has the largest mutual information distances across all values of epsilon, suggesting that it performs the worst under the metric of mutual information distance. Since Mode B generates synthetic data with independent attributes, there are no relationships modeled between the attributes. As a result, while individual attribute distributions might be preserved, the mutual information between pairs of attributes significantly deviates from that of the real dataset. This results in high mutual information distance values, particularly at lower epsilon values, where privacy preservation adds substantial noise and further distorts these inter-attribute relationships.

Mode C (Correlated Attribute Mode, k=1): Mode C performs better than Mode B but still shows relatively high mutual information fistance compared to Mode D. The introduction of weak correlations between attributes (with k=1) improves the preservation of inter-attribute relationships, but it is still not sufficient to closely match the real data's structure. While the synthetic datasets in Mode C show some improvement as epsilon increases, they still exhibit noticeable deviations from the real dataset's pairwise attribute relationships. At higher epsilon values, as the noise decreases, Mode C's synthetic datasets start to replicate the real data more closely, but the distance is still larger than in Mode D.

Mode D (Correlated Attribute Mode, k=2): Mode D performs the best in terms of mutual information distance. The introduction of stronger correlations (k=2) between attributes results in the closest replication of the real data's pairwise relationships across all epsilon values. Even at lower epsilon values, Mode D is able to preserve the dependencies between attributes better than Mode B and Mode C, leading to a lower mutual information distance. As epsilon increases, the noise decreases, and the synthetic datasets more closely match the real data's inter-attribute relationships, resulting in the lowest distances among the three modes.

Thus, Mode D (Correlated Attribute, k=2) achieves the best performance in preserving the mutual information between pairs of attributes, followed by Mode C (Correlated Attribute, k=1), with Mode B (Independent Attribute) showing the largest deviations from the real data due to its failure to capture inter-attribute dependencies. The plots confirm that stronger attribute correlations lead to better replication of the real dataset's structure, but independent attributes struggle to capture these relationships effectively.



The box-and-whiskers plots above illustrate the aggregated difference in pairwise mutual information between the synthetic datasets generated under Modes B (Independent Attribute), C (Correlated Attribute, k=1), and D (Correlated Attribute, k=2) compared to the real dataset (hw_compas) at various epsilon values, ranging from 0.0001 to 100. The X-axis represents epsilon, which controls the privacy budget, while the Y-axis represents the aggregated difference in pairwise mutual information. Lower values of mutual information difference indicate that the synthetic dataset closely matches the real dataset in terms of attribute correlations and dependencies, while higher values reflect larger discrepancies between the synthetic and real data.

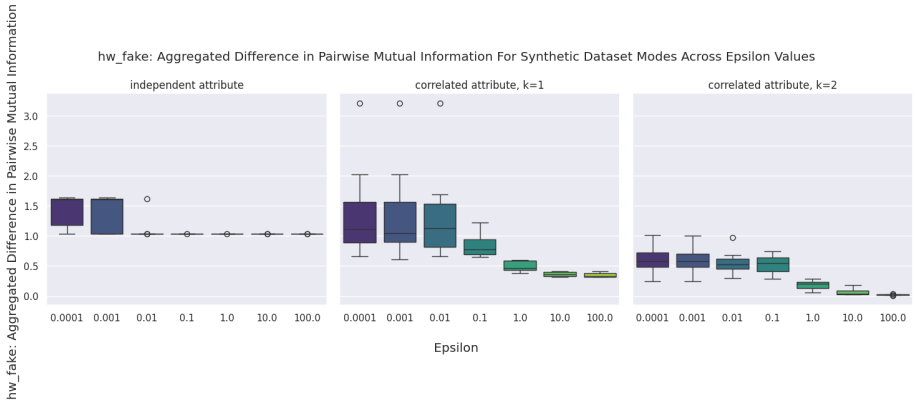
Mode B shows the smallest differences in mutual information, particularly at higher epsilon values (0.1, 1, 10, 100), suggesting that as the privacy budget increases, the synthetic datasets closely approximate the real dataset. Even at very low epsilon values (0.0001, 0.001), the difference in pairwise mutual information is still low, but there is a slight deviation compared to the real dataset. As epsilon increases, these differences shrink further, indicating that Mode B's synthetic data better matches the mutual information properties of the real data with less noise.

For Mode C, the aggregated difference in mutual information is higher than Mode B across all epsilon values, especially at smaller epsilon values such as 0.0001, 0.001, and 0.01. As epsilon increases, the difference decreases, showing that the synthetic datasets become more similar to the real dataset, similar to Mode B. However, the weak correlations (k=1) introduced in Mode C do not replicate the real data's pairwise mutual information as well as Mode B, particularly at higher epsilon values.

Mode D shows a similar trend to Mode C, but with a greater reduction in

the mutual information difference as epsilon increases. The stronger correlations in Mode D (k=2) allow the synthetic data to more closely match the real dataset's structure. However, at lower epsilon values, Mode D exhibits higher variability in the mutual information distance, indicating that stronger privacy protection (lower epsilon) results in more noise and greater deviations from the real data. As epsilon increases, Mode D's synthetic datasets better align with the real data's mutual information, though some variability remains at higher epsilon values.

In conclusion, Mode B achieves the closest match to the real dataset in terms of pairwise mutual information, with the smallest aggregated differences across all epsilon values. Modes C and D show higher differences in mutual information due to the varying strength of correlations between attributes. While Mode D (with stronger correlations) reduces the gap compared to Mode C, Mode B consistently demonstrates the most accurate replication of the real data's mutual information, particularly when epsilon values are smaller. Overall, the results suggest that Mode B is the most effective at preserving the mutual information properties of the real dataset, especially at lower epsilon values.



The box-and-whiskers plots above illustrate the aggregated difference in pairwise mutual information between the synthetic datasets generated under Modes B (Independent Attribute), C (Correlated Attribute, k=1), and D (Correlated Attribute, k=2) compared to the hw_fake dataset at various epsilon values, ranging from 0.0001 to 100. The X-axis represents epsilon, which controls the privacy budget, while the Y-axis represents the aggregated difference in pairwise mutual information. Lower values of mutual information difference indicate that the synthetic dataset closely matches the real dataset in terms of attribute

correlations and dependencies, while higher values reflect larger discrepancies between the synthetic and real data.

In Mode B, we observe the relatively high differences in mutual information at very low epsilon values (0.0001 and 0.001), suggesting that the privacy protection at these epsilon values introduces significant noise, causing larger discrepancies between the synthetic and original dataset (hw_fake). However, as epsilon increases (0.01 and beyond), the differences decrease considerably, and Mode B achieves the smallest differences in pairwise mutual information, especially at higher epsilon values (0.01, 0.1, 1, 10, and 100). This indicates that with weaker privacy protection, Mode B successfully generates synthetic datasets that very closely match the original data's pairwise mutual information properties. Despite the initial high discrepancies at low epsilon values, Mode B performs the best overall at higher epsilon values, where the privacy budget is larger, and less noise is introduced.

Mode C shows a higher aggregated difference in pairwise mutual information compared to Mode B across all epsilon values, particularly at lower epsilon values (0.0001, 0.001, and 0.01). This suggests that while Mode C introduces some correlation between attributes (k=1), it still generates datasets that are less similar to the original dataset's attribute dependencies compared to Mode B. As epsilon increases, the difference decreases, indicating that the synthetic data becomes closer to the real dataset, but the overall differences in mutual information remain higher than in Mode B. This suggests that while the correlations in Mode C improve the alignment with the original dataset (hw_fake), they are still insufficient to match the pairwise mutual information properties as well as Mode B.

Mode D shows the least change in aggregated difference across the epsilon values. The differences remain consistently lower compared to Mode C, indicating that Mode D's stronger correlations (k=2) make the synthetic data more stable and aligned with the real dataset. Mode D's differences are smaller than Mode C across all epsilon values, and the trend shows a relatively steady decrease in the differences as epsilon increases. This behavior suggests that Mode D has the advantage of stronger attribute correlations, which allow it to maintain a more consistent match with the real data's pairwise mutual information across the epsilon range. However, Mode D does not achieve the lowest differences at high epsilon values when compared to Mode B. While Mode D is better than Mode C, Mode B still performs better at higher epsilon values, with the smallest differences in mutual information.

In summary, Mode B exhibits the smallest aggregated differences in pairwise mutual information at higher epsilon values, indicating that it provides the best privacy-to-utility tradeoff, especially when the privacy budget is larger. Mode D performs better than Mode C in maintaining a consistent match with the original dataset's mutual information properties, but it does not perform as well as

Mode B at higher epsilon values. Mode C, despite its correlations (k=1), exhibits the highest differences, particularly at lower epsilon values, showing that it is not as effective at replicating the real data's pairwise mutual information properties. Therefore, while Mode B is the most accurate overall at higher epsilon values, Mode D provides more stable results across all epsilon values but with slightly less accuracy in preserving the original data's mutual information at higher privacy levels.

When comparing the aggregated difference in pairwise mutual information for hw_fake to the previous plots above for hw_compas, we can see some subtle differences in how the synthetic datasets replicate the original datasets' attribute relationships for hw_compas and hw_fake. For sets of plots (both hw_compas and hw_fake), Mode B (Independent Attribute Mode) performs the best overall, particularly at higher epsilon values, where it shows the smallest differences in mutual information. However, Mode B for hw_fake exhibits higher differences at low epsilon values (0.0001, 0.001) compared to the hw_compas plot, indicating that the synthetic data for hw_fake is less similar to the real data at these lower privacy levels. This suggests that hw_fake is more sensitive to noise introduced by differential privacy at lower epsilon values.

On the other hand, Mode D (Correlated Attribute Mode, k=2) maintains lower differences in pairwise mutual information across the epsilon range for hw_fake compared to hw_compas. This shows that Mode D's stronger correlations (k=2) allow for more stability in the synthetic data's alignment with the original dataset hw_fake, regardless of the privacy budget, but it does not outperform Mode B at higher epsilon values. Mode C (Correlated Attribute Mode, k=1) performs worse than Mode B and Mode D in both sets of plots for hw_compas and hw_fake, with higher differences in mutual information, especially at lower epsilon values.

Thus, while Mode B is still the most accurate overall in both datasets, the differences at low epsilon values for hw_fake suggest that the privacy-utility tradeoff may be more challenging to balance for synthetic datasets like hw_fake, requiring higher epsilon values for better data fidelity. In contrast, hw_compas shows that Mode B performs more consistently well across all epsilon values.