New York University **Erin Kong**
Responsible Data Science **Homework 3**
Prof. Umang Bhatt **April 27 2025**

Problem 1

Randomized Response

a. An individual flips a fair coin (probability 50/50), where tails leads to answering truthfully and heads leads to always answering "yes." This coin flip mechanism is differentially private with an epsilon value of $\epsilon = \ln(2) \approx 0.693$. To determine this, note the following probabilities.

   i. The probability of responding "yes" given the true answer is "no" is 0.5.

   ii. The probability of responding "yes" given the true answer is "yes" is 1.

   iii. According to the definition of differential privacy, the ratio between these probabilities is:
   $$\frac{1}{0.5} = 2$$
   , which means
   $$\epsilon = \ln(2) \approx 0.693$$

   Thus, this mechanism satisfies differential privacy with an $\epsilon = \ln(2)$. A smaller ϵ means stronger privacy with less information leakage. A larger ϵ has higher accuracy but with weaker privacy guarantees.

   The randomized response mechanism using a fair coin has an epsilon of 0.693. This is a considered a small epsilon, suggesting that the mechanism has a stronger privacy and less leakage of information, satisfying differential privacy.

b. The mechanism where an individual uses two fair six-sided dice, has a differential privacy parameter epsilon of approximately 1.386. To justify this, note the probabilities as follows:

   i. The probability of responding "yes" given the true answer is "yes" is:
   $$0.5 \times 1 + 0.5 \times \frac{1}{3} = \frac{2}{3}$$
   This accounts for a 50% chance of answering truthfully ("yes") and a 50% chance of answering "yes" because of the second dice, which has a probability $\frac{1}{3}$.

   ii. The probability of responding "yes" given the true answer is "no" is calculated as:
   $$0.5 \times 0 + 0.5 \times \frac{1}{3} = \frac{1}{6}$$
   There is a 50% chance of answering truthfully ("no") and a 50% chance of answering "yes" because of the second dice, which has a probability $\frac{1}{3}$.

iii. Using the definition of differential privacy, the ratio of these probabilities is:

$$\frac{\frac{2}{3}}{\frac{1}{6}} = 4$$

So, the privacy parameter is:

$$\epsilon = \ln(4) \approx 1.386$$

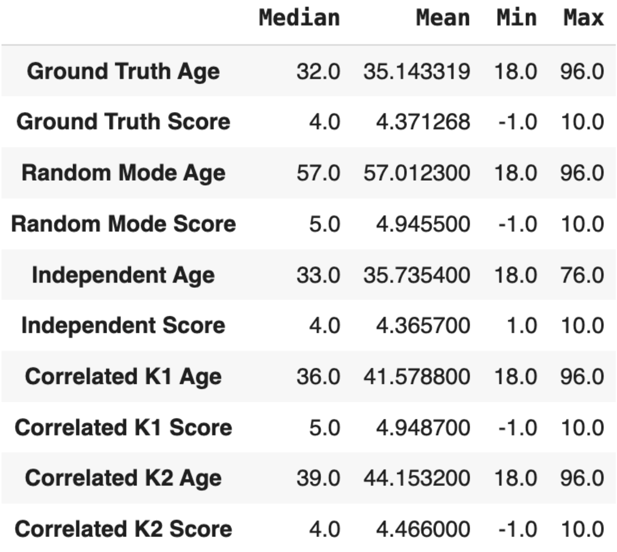
Therefore, this two-dice randomized response mechanism satisfies differential privacy with a privacy parameter of 1.386.

Problem 2

Privacy-Preserving Synthetic Data

(a) Queries on synthetic datasets and comparison on the corresponding real dataset.

Qi To do a comparison on the ground truth value from HW Compas, I have created a number of arrays consisting of aggregations of age and score. Then, I created and added these arrays to the respective columns in a pandas data frame. The random mode we used to generate our synthetic datasets didn't account that age is an integer between 18 and 96. The synthetic data in random mode for age revealed a minimum of 0 and maximum of 100, which wasn't in range. After I set the minimum and maximum of the random generator to the appropriate parameters, I was able to aggregate the age and score of all modes.



	Median	Mean	Min	Max
Ground Truth Age	32.0	35.143319	18.0	96.0
Ground Truth Score	4.0	4.371268	-1.0	10.0
Random Mode Age	57.0	57.012300	18.0	96.0
Random Mode Score	5.0	4.945500	-1.0	10.0
Independent Age	33.0	35.735400	18.0	76.0
Independent Score	4.0	4.365700	1.0	10.0
Correlated K1 Age	36.0	41.578800	18.0	96.0
Correlated K1 Score	5.0	4.948700	-1.0	10.0
Correlated K2 Age	39.0	44.153200	18.0	96.0
Correlated K2 Score	4.0	4.466000	-1.0	10.0

Figure 1: Ground Truth and Synthetic Dataset Comparison Table

The comparison between synthetic and the ground truth data sets show a significant difference in accuracy across all the modes. Random mode was observed to have the least accurate results, with a significant overestimation of both age and score averages, despite keeping the minimum and maximum, perhaps because it was set in the generator. Independent attribute was observed to have the highest accuracy in mean and median with the ground truth. However, looking closer at the maximum value of 76, it suggests that the range of values were smaller than the rest of the modes. In the Correlated K1 mode, it showed a better accuracy in overall structure than other modes, but included a higher bias in age and score. To add, correlated k2 mode furthered the existing bias, leading to a higher median and mean than correlated k1. Oddly, the accuracy of correlated k2 score was much better than correlated k1. It is noted that the independent mode had the most accurate scores in age and score, suggesting it is more reliable in individual features. Correlated modes can be more valuable for feature dependence, but require careful calibration, as existing bias can be amplified with this mode.

Qii Now, we will see how well random and independent attribute mode replicated the original distribution for age and sex.

The distributions of age and sex in the real and synthetic datasets generated under random and independent modes portrayed a disparity in replication, supported through distribution charts and statistical measures. In figure 2 below, the distribution for age shows that in random mode, it produces a uniformity rather than the real data's right skew distribution. To contrast, independent mode does better replicating the original data, but fall short in the frequency at older ages. In the distribution of sex, random mode was shown to have an almost even distribution between male and female, incorrectly replicating the real data of high males. On the other hand, independent mode shows a better distribution closely paralleling the shape of the real data.

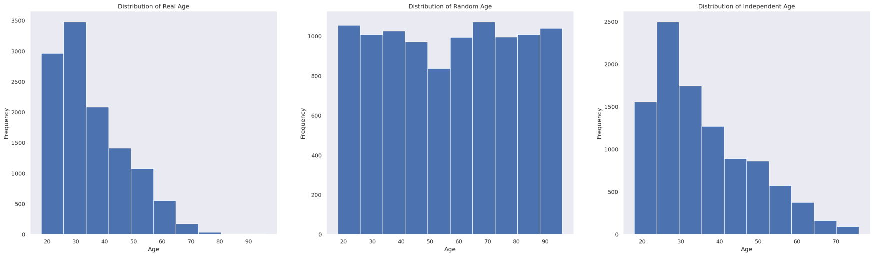


Figure 2: Ground Truth and Synthetic Dataset Distribution for Age

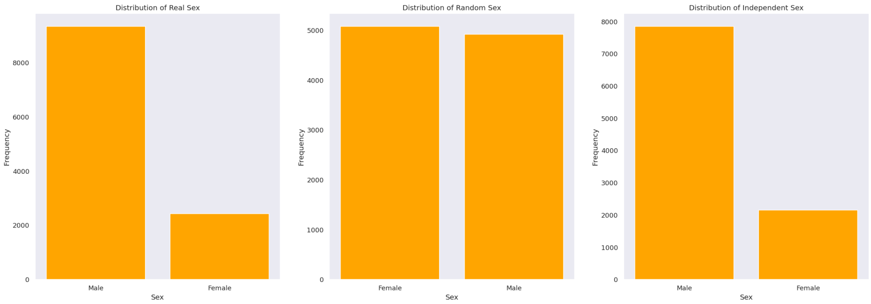
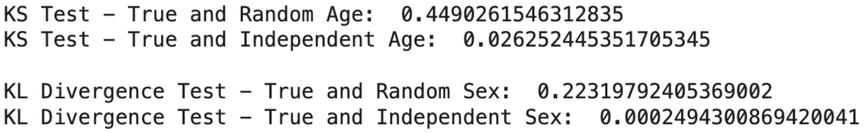


Figure 3: Ground Truth and Synthetic Dataset Distribution for Sex

To compute the real and synthetic datasets under random and independent modes using the Two-sample Kolmogorov-Smirnov (KS) test and Kullback-Leiber (KL) divergence test, it was applied to numerical and categorical attributes, respectively. In the KS-test, age in random mode had a value of 0.449, indicating a wide difference between the real and synthetic age distributions, supporting the visual distributions in earlier figures. The KS-test for independent attributes in age computed a comparably lower statistic of 0.026, suggesting that there is a similarity with the real data. For sex, the KL divergence test of random mode had an output of 0.223, supporting that there is still a deviation from the original dataset. Independent mode for sex had an output of 0.00025, achieving the best result of the two modes. The results of the visual distributions and KS-KL-tests supports that independent attribute mode is more accurate than random mode in both age and sex.

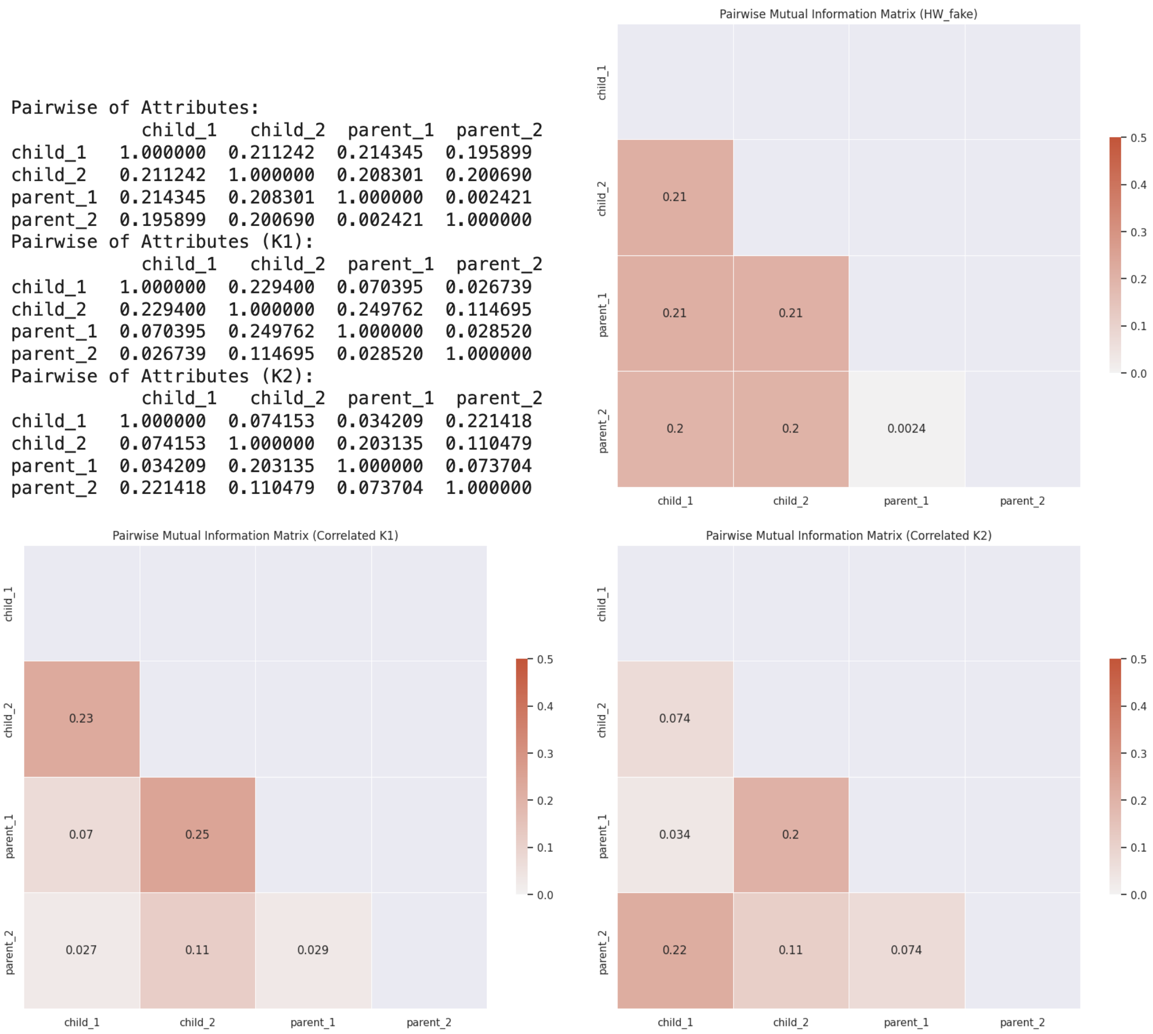


```
KS Test - True and Random Age:  0.4490261546312835
KS Test - True and Independent Age:  0.026252445351705345

KL Divergence Test - True and Random Sex:  0.22319792405369002
KL Divergence Test - True and Independent Sex:  0.0002494300869420041
```

Figure 4: KS and KL test results for age and sex

(b) To evaluate the correlated k=1 and k=2 attribute modes, the distance over pairwise correlation coefficients was computed, portraying visual differences in the results. In the original fake dataset, the mutual information of the child and parent attributes were all moderately around 0.2, excluding parent 1 and parent 2 at 0.0024. Under k=1, the correlation attribute's mutual information showed comparatively different results, with some relationships increasing to 0.23 (child 2 and parent 2)and 0.25 (child 2 and parent 1), and others decreasing to as low as 0.027 (child 1 and parent 2). In k=2, the mode's mutual information changed again, showing weaker relationships (child 1 and child 2) at 0.074 and stronger ones (child 1 and parent 2) at 0.22, interestingly swapping from the earlier k=1 correlation attribute. The accuracy of correlated attribute mode with k=1 proved a better preservation in structure in comparison to k=2.



```
Pairwise of Attributes:
            child_1   child_2  parent_1  parent_2
child_1   1.000000  0.211242  0.214345  0.195899
child_2   0.211242  1.000000  0.208301  0.200690
parent_1  0.214345  0.208301  1.000000  0.002421
parent_2  0.195899  0.200690  0.002421  1.000000
Pairwise of Attributes (K1):
            child_1   child_2  parent_1  parent_2
child_1   1.000000  0.229400  0.070395  0.026739
child_2   0.229400  1.000000  0.249762  0.114695
parent_1  0.070395  0.249762  1.000000  0.028520
parent_2  0.026739  0.114695  0.028520  1.000000
Pairwise of Attributes (K2):
            child_1   child_2  parent_1  parent_2
child_1   1.000000  0.074153  0.034209  0.221418
child_2   0.074153  1.000000  0.203135  0.110479
parent_1  0.034209  0.203135  1.000000  0.073704
parent_2  0.221418  0.110479  0.073704  1.000000
```

Pairwise Mutual Information Matrix (HW_fake)
child_1
child_2
parent_1
parent_2
0.21
0.21
0.21
0.2
0.2
0.0024
0.5
0.4
0.3
0.2
0.1
0.0
Pairwise Mutual Information Matrix (Correlated K1)
0.23
0.07
0.25
0.027
0.11
0.029
Pairwise Mutual Information Matrix (Correlated K2)
0.074
0.034
0.2
0.22
0.11
0.074

Figure 5: Pairwise Mutual Information Matrix Comparison

To begin this problem, the minimum and maximum were set at 18 and 96, respectively for the random mode. The epsilon for the independent mode was set to 0.01 while correlated was the same with a Bayes of k=1. By fixing a random seed and creating a for loop at range 10, I created 10 synthetic datasets of each mode, put the aggregates into a numpy array, and put it into a list. After this, I was able to have 3 dataframes for the aggregates of each mode and concatenate them into one larger dataframe with an added column for the mode type. Then, plotted the box and whisker plots for the median, mean, minimum, and maximum of the age attributes across the modes as shown below.

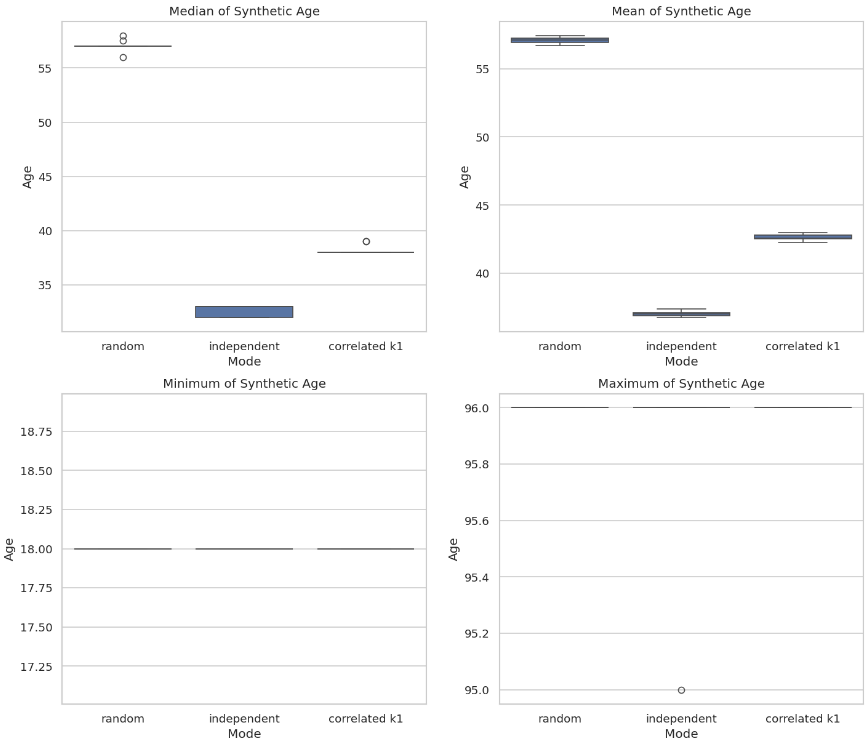


Figure 6: Variability of Synthetic Data

As shown in the figure above, the minimum and maximum of synthetic age stayed at the same level for all modes, with an exception to an outlier of the maximum in the independent mode. In the median of the synthetic datasets, the random mode had higher values than the independent and correlated k=1 attributes. To contrast, the shape of the distribution stayed similar to the median, where the random mode had considerably higher values than that of independent and correlated k=1 attributes. Now, we will compare this to the ground truth values from the real dataset.

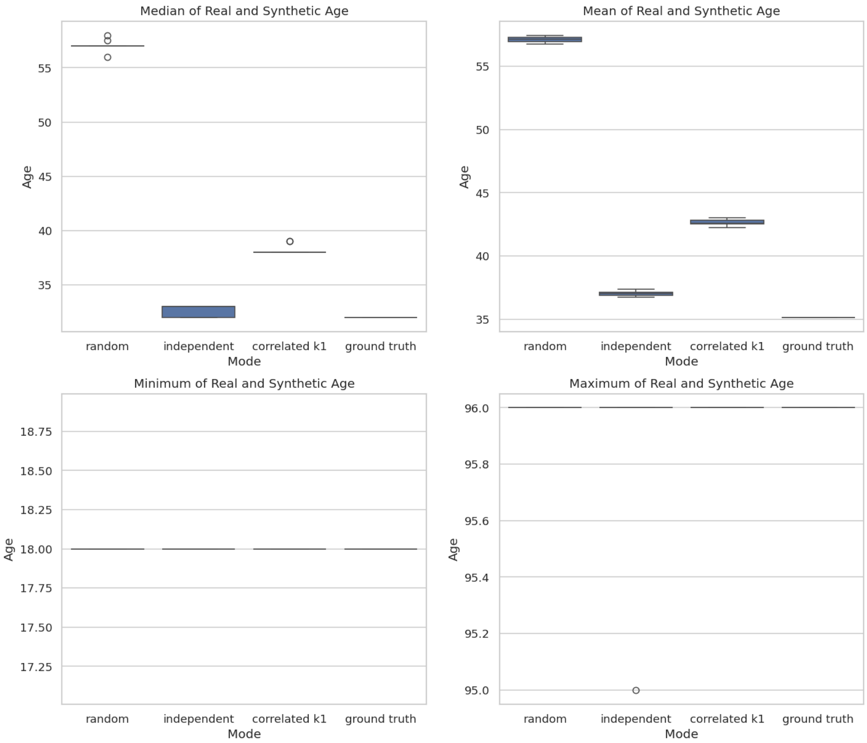


Figure 7: Variability of Real and Synthetic Data

The figure above shows the mean, median, min, and max for random mode, independent attribute, correlated k=1 attribute, as well as the ground truth values.

In the median subplot, the independent attribute resembled the closest to the ground truth, correlated k=1 was in close second, and the random was the furthest from the real value. This would assume that the independent attribute had the most accurate preservation of the real dataset. In the mean subplot, the independent attribute resembled the closest to the ground truth again, correlated k=1 attribute was in close second again, and the random was again the furthest from the real value. This would again assume that the independent attribute had the most accurate preservation of the real dataset. In the minimum subplot, there was no variability across modes, all producing a minimum of 18. In the maximum subplot, there was little variability across modes, all producing maximum of 96, with exception to an outlier in the independent attribute.

The independent attribute is observed to have the most accurate metrics compared to the ground truth of the real data. Because the mode works by modeling the features separately based on the original distribution, it would match the distribution of age and closely replicate the real dataset.

The random mode had the least accurate metrics compared to the ground truth of the real data. Because this mode works by drawing values from 18 to 96 uniformly, the median of the min and max values would be about the middle of the synthetic data distribution, which aligns closely to the behavior observed in figure 2 and figure 3.

The correlated attribute mode of k=1 had a higher bias, but a smaller variability across metrics. Because the mode works by preserving the relationships between attributes, it can exaggerate biases whilst maintaining the similar spread that the original ground truth, thus showing a slight shift in mean and median.

(c) Now, we will determine statistical properties of the data preserved as a function of the privacy budget.

i. On the sensitive attribute Race, epsilon = [0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1] was applied to the independent attribute, correlated attribute k=1, and correlated attribute k=2. I created a nested loop of the epsilon list to iterate 10 times in addition to the iterations of the epsilon values, and was able to create 3 Data Frames for each mode of 100 row values. For each synthetic dataset in every mode, I was able to compute the KL-divergence test between the real data and the synthetic data for each epsilon. With this, I plotted box and whisker plots with the epsilon and KL scores as my X and Y, respectively. The result is as follows:

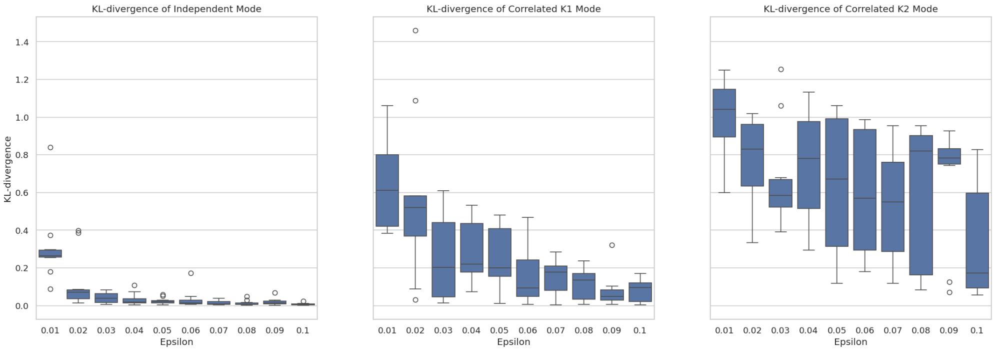


Figure 8: Race Distributions of KL-Divergence Scores

In the independent mode, the KL-divergence seemed to decrease as the epsilon increased, implying that the synthetic data approximates the real race distribution with even moderate privacy budgets. The variability for this mode was small, with an exception to an epsilon of 0.01 and 0.02.

In the correlated k=1 mode, the KL-divergence decreased as the epsilon increased. However, the variability is larger than that of the independent attribute, suggesting that the synthetic dataset is sensitive under differential privacy.

In the correlated k=2 mode, the KL-divergence had the largest values and worst performance. The KL-divergence values remained high despite being at large epsilon and the variability was substantial on all settings as well. This suggests that there are strong correlations and race isn't preserved well, perhaps with bias.

In the different generational settings, the independent attribute best preserves the race distribution. Preservation was observed to be the worst in correlated k=2, and better in correlated k=1, but equally as bad. As stronger correlations are applied, the lack of preservation is apparent in the variance of KL scores.

ii. To analyze how well the attribute relationships were preserved, we compared the pairwise mutual information of the synthetic datasets of the fake and real datasets. The epsilon used was [0.0001. 0.001, 0.01, 0.1, 1, 10, and 100] under the independent, correlated k=1, and correlated k=2 modes. Similar to the first section of (c), a nested loop of the epsilon were iterated 10 times to create 3 Data Frames for 70 row values (due to the smaller epsilon list). For every synthetic dataset, I created a function that took the absolute value of the mutual information difference, and aggregated across all pairs without repetition. With this, I plotted box and whisker plots with the epsilon and the mutual information difference as my X and Y, respectively. The result is as follows:

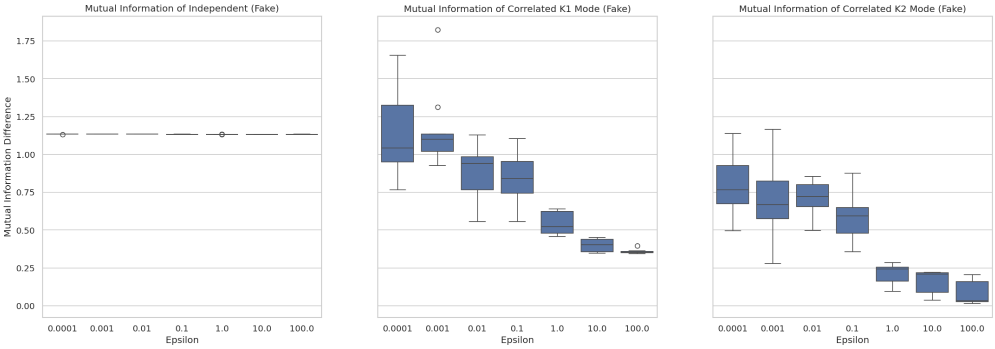


Figure 9: Mutual Information Difference of Fake Data

It is observed in the independent attribute that the mutual information difference of the fake data remained high in all of the epsilon values, which makes sense because all the features were sampled independently.

In the correlated k=1 and k=2 attribute, the mutual information differences were observed top decrease as the epsilon increased. For the epsilon values of 10 and 100, the correlated modes have low mutual information differences, which suggests that the relationships were preserved when there was less privacy. Correlated k=1 had a better preservation at a moderate privacy setting in comparison to correlated k=2.

The independent mode fails to preserve the relationships between attributes, whilst the correlated attributes improve as the epsilon increases.

The same process on the fake dataset above was done to the real dataset. The results are shown below:

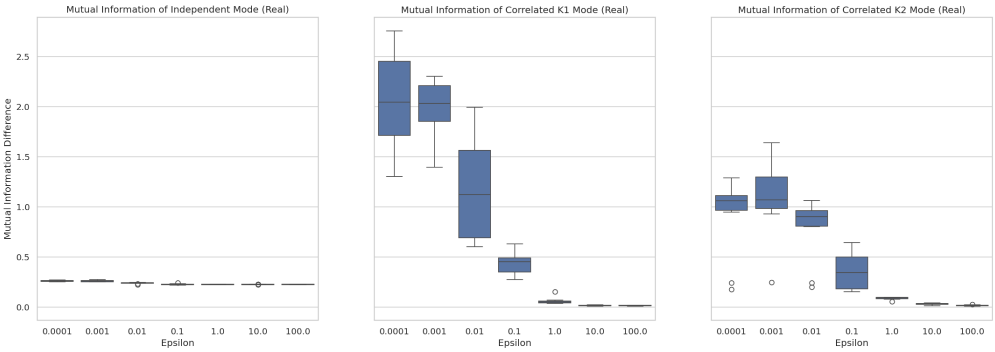


Figure 10: Mutual Information Difference of Real Data

The independent mode is observed to have a consistent high mutual information difference, suggesting that the mode does not maintain preservation in relationships between attributes.

The correlated k=1 and k=2 modes had a strong improvement of mutual information differences as the epsilon increased, similar to the fake dataset. However, despite similarities, the mutual information differences of the real datasets were overall slightly higher than that of the fake datasets. This suggests that the real datasets had more dependencies and relationships that were harder to preserve under the low epsilon.

In the preservation of mutual information, it is suggested in the data that using the correlated attributed methods depends highly on the privacy budget. As the privacy budget weakens, the preservation increases and vice versa. There exists an inherent tradeoff of privacy and accuracy in differential privacy. With a smaller epsilon, there is more privacy as there is additional noise, but becomes less like the real data. With a larger epsilon, there is less privacy and with that less noise (randomization), thus becoming more accurate to the real data as well.

Key comparisons between the real and fake datasets is as follows:

Mode	**Real Data**	**Fake Data**
Independent Attribute	MI difference is about uniform with limited variation.	MI difference is high and constant across all epsilon.
Correlated Attribute (k=1)	MI difference is larger at small epsilons but drops steeply as epsilon increases, reaching low values at large epsilons.	MI difference decreases gradually as epsilon increases, reaching approximately 0.01 at large epsilons.
Correlated Attribute (k=2)	MI difference starts large and gradually decreases as epsilon increases.	MI difference starts large but drops steadily as epsilon increases.
Overall Preservation	Real data contains strong attribute dependencies; low epsilon causes larger MI differences.	Fake data contains weak attribute dependencies; smaller MI differences overall.

Table 1: Comparison of Mutual Information Differences in Real and Fake Data.