Analyzing Racial Disparities

Preethi Narayan

Apr 13, 2022

Problem 1

Part A

(5 points) Give three distinct reasons why racial disparities might arise in the predictions of such a system.

A. **Technical bias** - Technical bias may cause racial disparities in the data. Technical bias may occur as a result of different weights being assigned to certain races / neighborhoods within the data. The system may have attached heavier likelihood of crime to certain areas/races, which would result in the system assigning more people ot patrol those areas more heavily. This is an example of technical bias that may exist in the model.

B. **Historical bias** - From our previous readings and from other sources, it is very apparent that historically, police officers have taken race into consideration when making decisions on which people should be searched/detained, or which neighborhoods should be patrolled. This affects training data because it means that this bias will appear in past data archives, thereby affecting the training data. As a result, the training data will carry bias from the past. This previous bias being incorporated into the model will result in further perpetuation of the system that allows minority neighborhoods to continue to be overpatrolled, while white neighborhoods are under patrolled.

C. **Feedback loop** - A feedback loop results from what happens due to the model's first recommendations. If a model recommends that police should monitor a particular low-income or minority filled neighborhood due to some other bias (possibly historical bias), then police will be sent to that area and they will make arrests. Due to those arrests having been made, more police will be sent to the area and more arrests will be made on top of the previous ones. This will result in a constant positive feedback loop of more police being sent to these underprivileged neighborhoods, while less are being sent to privileged neighborhoods and white neighborhoods. The system will learn over time and

assume that it is correct in sending less police to white neighborhoods and more to underprivileged neighborhoods, and will continue to do so.

Part B

(5 points) Propose two mitigation strategies to counteract racial disparities in the predictions of such a system. Note: It is insufficient to state that we could use a specific pre-, in- or post-processing technique that we covered in class when we discussed fairness in classification. Additional details are needed to demonstrate your understanding of how the ideas from fairness in classification would translate to this scenario.

A. **Synthetic data** - One mitigation strategy in this situation could be to use synthetic data. If synthetic data without any previous biases is used, then the model won't carry any historical bias into the model. Using synthetic data would reduce the location bias that is held by the model. Considering that this data itself might have bias as it is synthetically generated and isn't using true historical data, we could use a synthetical to create hypothetical neighborhoods that are under policed to off set the bias. The data can be split into train and test sets as normal, but then the test sets can be used to train even more models, and this can be done using the three modes that we have previously discussed. Then we can continue to tweak the epsilon value to ensure that the privacy preserved in the synthetic dataset is controlled. Then, fairness metrics can be calculated on all three modes, and the accuracy can be compared to that of the test data.
B. **Reweighting** - The training data could be reweighted to ensure that additional weightage is not being added to underprivileged neighborhoods. The weights could be distributed so that some of the weights shift from underprivileged neighborhoods and towards ehite neighborhoods so that additional and unnecessary patrolling will not be sent towards the underprivileged neighborhoods due to the earlier weights of the training data. As the reweighting algorithm takes frequency counts into account while reweighting, the data moving forward will assign higher weights to regularly under policed neighborhoods, also referred to as those with low frequency counts, and the data will assign lower weights to regularly overpoliced neighborhoods, also referred to as those with high frequency counts. Moving forward, the data can be pre processed with the reweighting algorithm, and then any classifier models can be trained on the re-weighted data, and this should mitigate bias.

Problem 2

Part A

(15 points) The simplest version of randomized response involves flipping a single fair coin (50% probability of heads and 50% probability of tails). Suppose an individual is asked a potentially incriminating question, and flips a coin before answering. If the coin comes up tails,

he answers truthfully, otherwise he answers “yes”. Is this mechanism differentially private? If so, what epsilon value does it achieve? Carefully justify your answer.

A. Yes, this mechanism is differentially private. According to our reading, “Achieving differential privacy revolves around hiding the presence or absence of a single individual”. In the hypothetical situation above, one will never know whether or not the person answering the question is telling the truth ot not. If the individual says yes, the analyst in question will never know whether they are answering yes because they flipped heads, or if they are telling the truth. As a result, this mechanism is differentially private.

Problem 3

Part A

Q1 (hw_compas only): Execute basic statistical queries over synthetic datasets. The hw_compas has numerical attributes age and score. Calculate the median, mean, min, max of age and score for the synthetic datasets generated with settings A, B, C, and D (described above). Compare to the ground truth values, as computed over hw_compas. Present results in a table. Discuss the accuracy of the different methods in your report. Which methods are accurate and which are less accurate? If there are substantial differences in accuracy between methods - explain these differences.

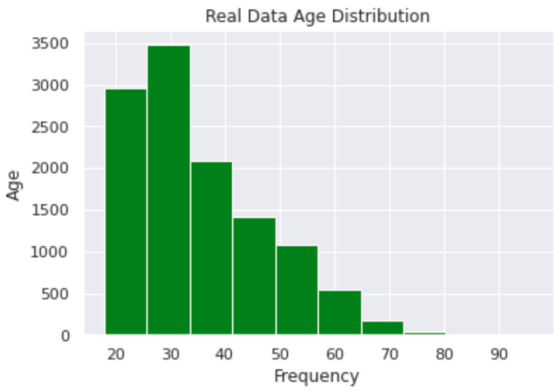
	truth age	truth score	random real age	random real score	independent real age	independent real score	correlated k1 age	correlated k1 score	correlated k2 age	correlated k2 score
Mean	35.143319	4.371268	50.1731	4.9392	35.7354	4.3657	41.5788	4.9487	44.1532	4.466
Min	18.000000	-1.000000	0.0000	-1.0000	18.0000	1.0000	18.0000	-1.0000	18.0000	-1.000
Max	96.000000	10.000000	100.0000	10.0000	76.0000	10.0000	96.0000	10.0000	96.0000	10.000
Median	32.000000	4.000000	51.0000	5.0000	33.0000	4.0000	36.0000	5.0000	39.0000	4.000

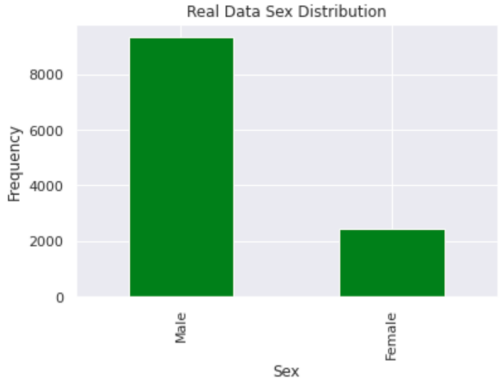
There were some substantial differences in the accuracies. One significant difference that I notices was the score and attribute differences between that of the random real age and the random real score from the truth age and the truth score. This occurred because randomly generated data focused less on generating data that follows the statistical patterns within the original dataset, and rather focused on creating data that was a similar shape and size.

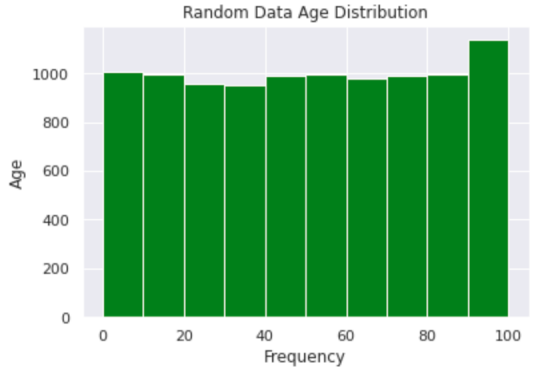
The mean and median of the correlated k1 and k2 values were higher than that of the original mean and medians. The scores were close to the original. In the place of a target protected feature, this dataset has values that are instead created from a conditional distribution of the existing values in the data.

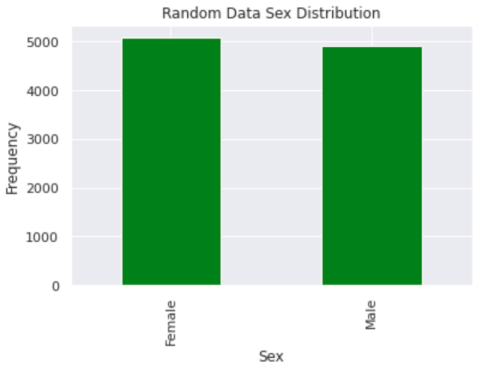
The independent attribute is the most accurate. The age and score were very close to the original values. In the place of a target protected feature, this dataset has a values that follow a similar distribution to the original values.

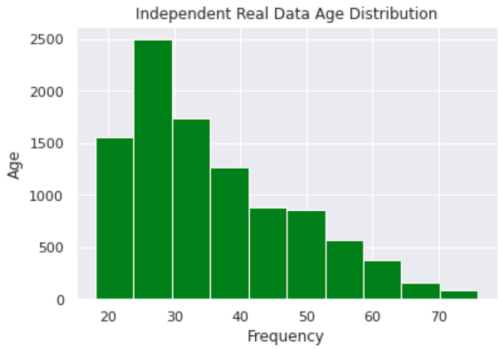
Q2 (hw_compas only): Compare how well random mode (A) and independent attribute mode (B) replicate the original distribution. Plot the distributions of values of age and sex attributes in hw_compas and in synthetic datasets generated under settings A and B. Compare the histograms visually and explain the results in your report. Next, compute cumulative measures that quantify the difference between the probability distributions over age and sex in hw_compas vs. in privacy-preserving synthetic data. To do so, use the Two-sample Kolmogorov-Smirnov test (KS test) for the numerical attribute and Kullback-Leibler divergence (KL-divergence) for the categorical attribute, using provided functions ks_test and kl_test. Discuss the relative difference in performance under A and B in your report.

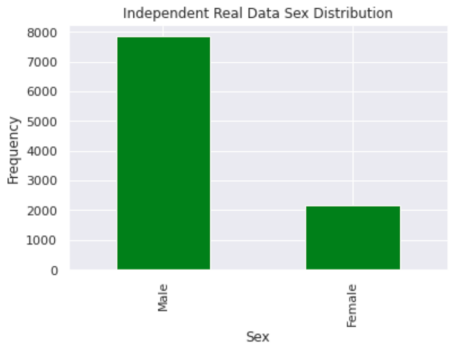












The distributions of the real data and the independent data are very similar. The Age is skewed to the right and there are similar ratios of males to females. This occurs because of how the mode calculates the values, the values are calculated by estimation of the probability distributions based on frequency. As mentioned earlier, the random mode does not generate values that follow the source's statistical probability patterns, and as a result the data would follow a different distribution. The random mode causes the data to have an even distribution of almost everything, leading age and sex to be distributed almost equally.

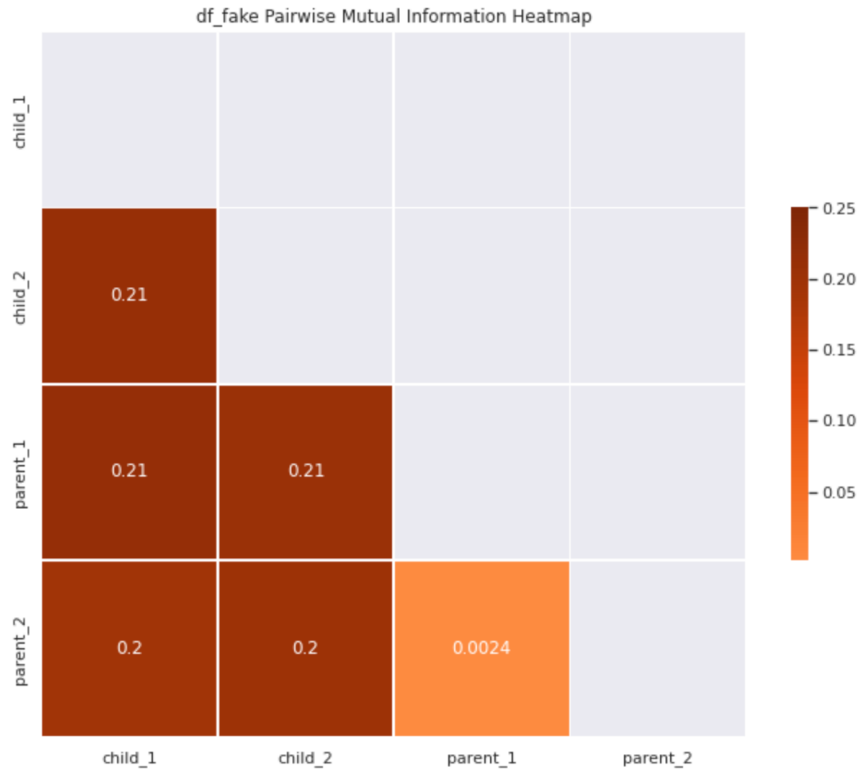
```
KS test for age between Truth and Independent:  0.026252445351705345
KL test for age between Truth and Independent:  0.0002494300869420041

KS test for age between Truth and Random:  0.3735091775112699
KL test for age between Truth and Random:  0.22319792405369002
```

A higher KL divergence value means that we have not matched the true distribution that well with our approximation. Lower the KL divergence value, the better a match there is between the true distribution of the data and the approximation.This means that the distributions between the Truth and Independent values were much closer than that of the Truth and Random values.

Q3 (hw_fake only): Compare the accuracy of correlated attribute mode with k=1 (C) and with k=2 (D). Display the pairwise mutual information matrix by heatmaps, showing mutual information between all pairs of attributes, in hw_fake and in two synthetic datasets (generated under C and D). Discuss your observations in your report, noting how well / how badly mutual information is preserved in synthetic data.

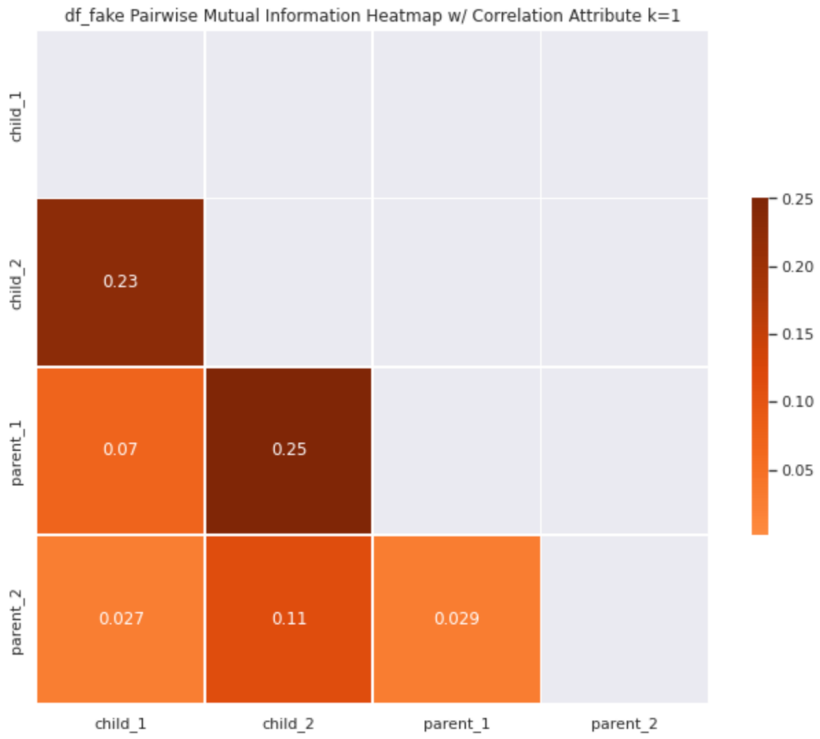
Pairwise:

	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



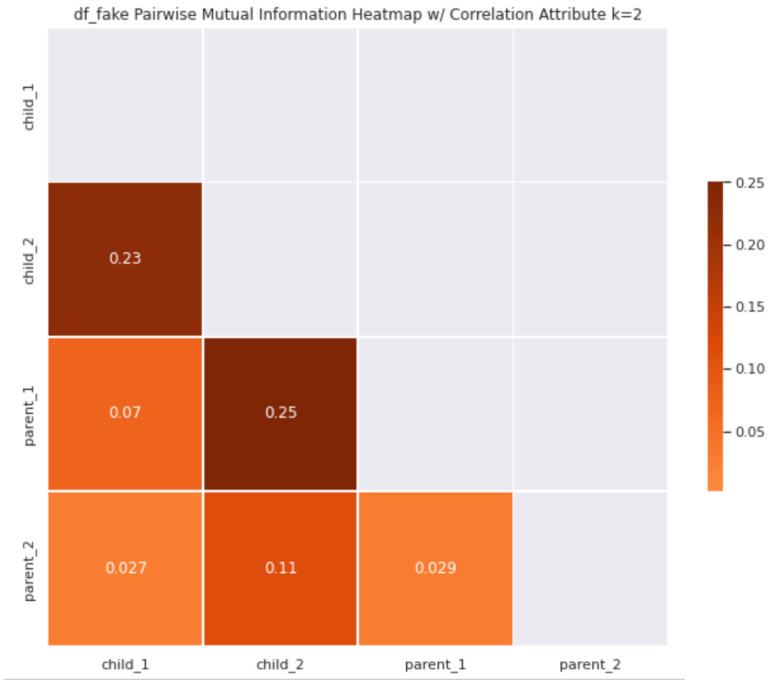
Pairwise w/ Correlation k=1:

	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 w/ Correlation k=2:

	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



Child1-Child2: The setting with correlated attribute mode with k=1 preserved more in this case than the setting with k=2. The preservation was similar overall.

Parent2-Parent1: The setting with correlated attribute mode with k=1 (C) preserves the most in this case. Both settings C and D preserve more than that of df_fake. Though neither C nor D greatly follow the pattern of df_fake, setting D is more similar to it than C.

Parent2-Child2: C preserves very little in this scenario. While D and df_fake are the most similar in this situation, df_fake preserves the most in this scenario.

Parent2-Child1: df_fake preserves the most in this scenario.

Parent1-Child2: df_fake in this case preserves more than both settings C and D. Neither setting resembles df_fake. Setting D is still most similar to df_fake.

Parent1-Child1: Setting D preserves the most in this scenario. Neither setting resembles df_fake. D is the closest to df_fake.

Part B

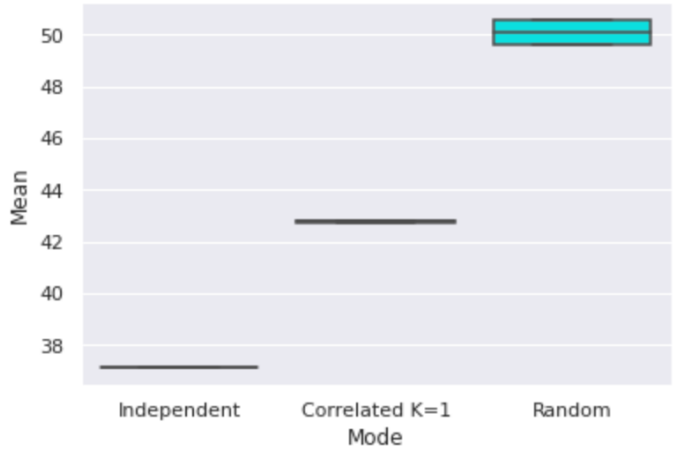
(10 points, hw_compas only): Study the variability in the mean and median of age for synthetic datasets generated under settings A, B, and C.

To do this, fix epsilon = 0.1, and generate 10 synthetic datasets (by specifying different seeds) for each setting A, B, and C. Calculate the mean and median of age in each of the generated datasets.

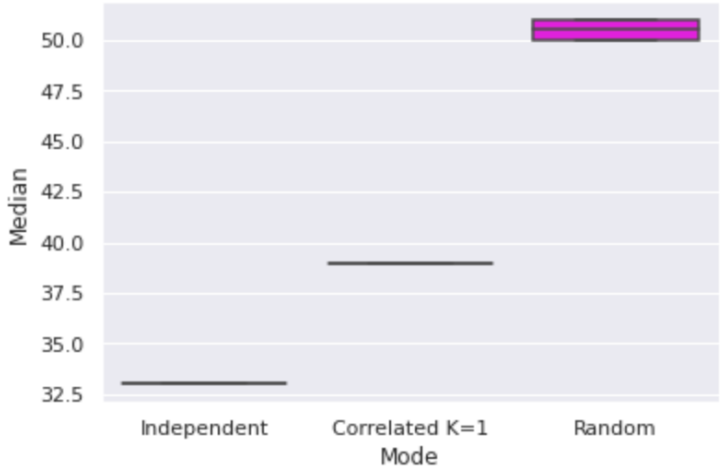
Then, for each setting, plot the 10 median values and the 10 mean values using a box-and-whiskers plot. Compare these metrics to the ground truth median and mean from the real data. Carefully explain your observations: which mode gives more accurate results and why? In which cases do we see more or less variability?

Specifically, you should generate 30 datasets in total: 10 under setting A, 10 under setting B, 10 under setting C. For the box-and-whiskers plots, we expect to see two subplots: one for the mean and one for the median, with the three settings (A, B and C) along the X-axis and age on the Y-axis. You should include these plots in your report.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6c8f6f490>
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6c98f37d0>
```



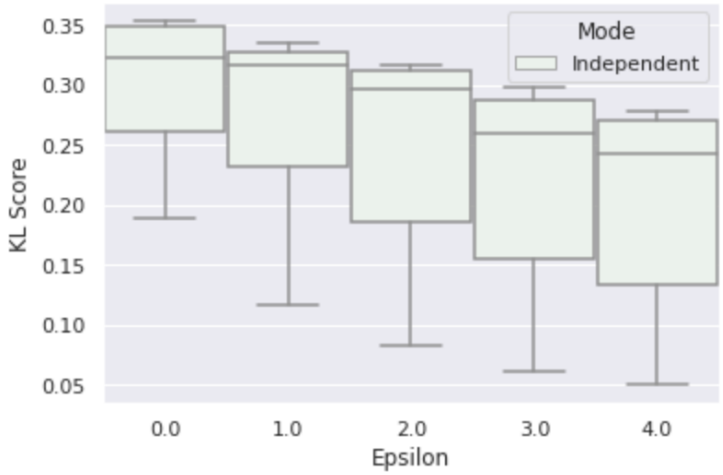
As seen in these boxplots, the mean and median have the most variability when they are generated in random mode. When it comes to the independent and the correlated k=1 modes, the variability is more or less the same. The independent mode gives the most accurate results and the correlated mode gives the second most accurate results. In regards to the minimum and

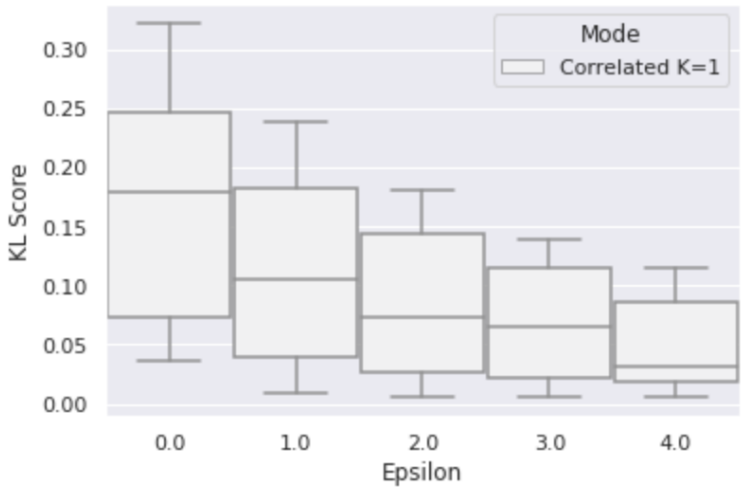
maximum values, the independent and correlated modes has relatively the same minum and maximums. The independent mode performed the best, further proving our earlier results when we found that it preserves the statistical patterns and distributions of the original data.

Part C

(10 points, hw_compas only): Study how well statistical properties of the data are preserved as a function of the privacy budget, epsilon. To see robust results, execute your experiment with 10 different synthetic datasets (with different seeds) for each value of epsilon, for each data generation setting (B, C, and D). Specifically, you should: Compute the KL-divergence over the attribute race in hw_compas. For each setting (B, C, and D), vary epsilon from 0.02 to 0.1 in increments of 0.02. Specifically, the epsilons are [0.02, 0.04, 0.06, 0.08, 1]. In total, you should generate 150 synthetic datasets (3*10*5) and calculate the KL-divergence for race in each dataset. Create three box-and-whiskers plots, one for each setting (B, C, D). Each plot should have epsilon on the X-axis and KL-divergence on the Y-axis. Discuss your findings in the report and include your plots.

The box plots that I generated have been pasted below. As one can clearly see in the box plots that were generated, as the epsilon increased, the KL score decreased. As I have mentioned earlier, lower the KL divergence value, the better a match there is between the true distribution of the data and the approximation. Therefore, as the epsilon decreased, the approximations became better. This happened because a lower epsilon value indicates that there exists higher privacy preserved in the data. A low epsilon value indicates that the synthetic and real datasets are extremely similar. A high epsilon value indicates that the synthetic and real datasets are not similar. Lower epsilon values indicate that there is less noise in the data, and they closely resemble the original values. The two datasets that were most similar to the original dataset were the independent mode data and the correlated mode data which had k = 1.






```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6c989b090>
```