CS 2390: HW 3

Question 1: There are 41 main and 337 sub-genres in music. Analysis reduces these to seven relevant genres which is already a pretty narrow window. Using a statistical approach, I observe that Pop is the most popular, accounting for one-third of preferences across age groups. Therefore, Pop may likely reflect a TA's musical taste. For Kinan, who is in his late 20s and has not opted out of answering, Metal emerges as the probable favorite.

Question 2: Sotwe.com has the answer. He indeed is a metal lover. They are consistent with the answer from Q1. He started college in 2012(from LinkedIn), was born in 1994(from Facebook) and thus he is 29.

Q3: It is black. It was pretty easy. There are only 2 29 year olds. I'm using previous information and making a biased prediction that his favorite color is likely black rather than yellow(which is the favorite color of the other 29-year-old) as he is a metal lover :).

Q4: He falls in the 25 or more group. 4 different sports are mentioned in this group. Since he is from Syria(the most popular sport in Syria is soccer, just like Turkey, and their national team has made it to World Cup qualifiers), and since it also is the category with the most people, Kinan's favorite sport is likely Soccer.

Q5: Looking at the result of the query from last year, baseball gets eliminated as it is not included in the last year's results. This isn't as strong of a statement but based on your statement that his favorite sport didn't change, I'll also eliminate the e-sports option as it isn't the same text. Finally, I do the query with the exact age and as there is only one 29-year-old, it confirms that Kinan's favorite sport is soccer.

Q6: A larger ε makes the scale (b) of the Laplace distribution smaller. As a result, less noise is added to each count in the histogram, making the data more accurate as it is closer to the real values. This means weaker privacy guarantees as our queries stay relatively more informative, making us still able to identify individuals. The exact opposite thing happens with smaller ε -> larger scale -> more noise -> less accurate counts -> stronger privacy guarantee.

Q7: The peak appears around 1, thus the most likely value with more than 20% frequency is 1. Due to the symmetry of the Laplace distribution, it is likely that the mean is close to the actual value, thus the expected value is also around 1. With our epsilon value, we expect the noise to create sufficient variability around the true value. Although individual observations may vary widely, the expected value should still be near the true value due to the symmetry of the Laplace distribution. When the epsilon value is lower, with more noise added, the plot is flatter and

wider, leading to more uncertainty about the true count. On the other hand, when epsilon is higher, with less noise added, the plot is more peaked and narrower.

Q8: The exposed averages closely estimate the true average ages. We know Kinan's age and have some idea about his programming experience. We have a general idea of the distribution of the ages. The group Kinan most likely falls into has only a few members. Having some idea of the distribution for an average of 28 that group probably includes Kinan. Then this information can be used to deduct that he has more than 10 years of programming experience. This also adds up with the year he started college.
The group that Kinan probably belongs to only has 4(small number) members. When this is the case, the average is impacted more by the individual ages, leading to a more accurate guess. Thus, I'm confident about what I deducted. Some scenarios where inferences might be wrong are when some level of noise still impacts the accuracy of the exposed averages. Or they may be individual variabilities such as being an outlier. There may be misleading averages, especially if the distribution of ages within a group is skewed. If the TA's experience level falls into a large group, it would make specific inferences about the TA less reliable.

Q9: Things from Q8 still hold. Especially this portion: "The group that Kinan probably belongs to only has 4(small number) members. When this is the case, the average is impacted more by the individual ages, leading to a more accurate guess. " In summary, he is 29, he is in the group with 10 or more years of programming experience, and the average of that group of 4 is 28. Thus, I'm confident in my answers.

Q10: The budget is tracked on the client side. The client can simply modify the code or bypass the budget checks. The budget state isn't stored persistently. If you restart the application, the budget information will also reset or get lost. If multiple users are accessing the same data, there is no global limitation to the amount of queries that can be done on it.

Design for a Robust Enforcement Mechanism.
The budget tracking should be done on the server side. This way, each query request would only be permitted if there is still a budget. Persistent storage of the state of the privacy budget for consistency. Implement a centralized budget management system. That tracks and updates the privacy budget for each dataset. The alert mechanism for when the budget is running low. Implement rate limiting for API calls to prevent rapid consumption of the privacy budget