Project-1: Anonymity

In this report, we present the anonymization of flight data and the evaluation of utility in the context of *k*-anonymity. The project involved the use of a flight dataset, and we explored different values of *k*, defined our Quasi-Identifier (QID), measured utility, and demonstrated the impact of varying *k*-values on utility.

1. Values of *k* Used

We considered three different values of *k*: 2, 5, and 50. These values represent varying levels of anonymity and allowed us to assess the trade-off between privacy and data utility.

2. Quasi-Identifier (QID)

For this project, we defined our QID attributes as follows:

- **Gender**: This attribute is sensitive and can identify individuals. It is an essential component of the QID.
- **Airport Continent**: The airport continent can also contribute to identification, especially in the context of less common continents.
- **Age**: Age is considered as a QID attribute in the code. It can be sensitive in certain cases and was included in the QID to ensure comprehensive privacy protection.
- **Departure Date**: Although not as directly identifiable as some other attributes, the departure date can still be sensitive in some cases, especially when combined with other attributes.

We chose these attributes based on the potential for individual identification and the need to balance privacy with data utility. Gender, airport continent, age, and departure date were included to account for potential indirect identification, ensuring a robust approach to privacy preservation.

3. Utility Measurement

We assessed utility using a simple scoring system, where we started with a utility value of 1 and deducted points for various anonymization steps. The utility loss factors considered were:

- **Suppressing Attributes**: Each suppressed attribute resulted in a 1% reduction in utility.
- **Level 1 Generalization (Date)**: Generalizing departure dates to the month level resulted in a 5% reduction in utility.
- **Level 2 Generalization (Date and Age)**: Calculated the difference in months for the ‘Departure Date’ and the difference in ages for ‘Age’ within each group of records that are being generalized. These differences are scaled by a factor (1700) to ensure they contribute appropriately to the utility loss. Each month difference contributes a 0.001 reduction in utility, and each age difference contributes a 0.0001 reduction in utility.

4. Impact of *k* on Utility

- For *k* = 2, the utility score was the lowest, indicating the highest level of privacy but potentially limited utility for detailed analysis.
- For *k* = 5, the utility score was moderate, balancing privacy and utility.
- For *k* = 50, the utility score was the highest, indicating the lowest level of privacy but preserving the most data utility.