K-Anonymity and L-Diversity Implementation in Python:

Libraries Used:

- pandas: For data manipulation and analysis.
- numpy: For numerical operations.

Functions:

isKAnonymized:

- Checks K-anonymity and L-diversity conditions in a partition.

get_spans:

- Calculates spans of each column in a partition.

split:

- Splits a partition based on median or all categories.
- For numerical columns, we take the median of all the values and split the dataset into half based on median.
- For categorical column, we split dataset to n partitions, where n is number of categories in categorical column.

partition_dataset:

- Partitions dataset based on K-anonymity and L-diversity.
- Iteratively splits based on categorical and numerical columns.
- Merges categories not satisfying K-anonymity.
- Returns partitions satisfying conditions.

get_anonymize_dataset:

- Creates anonymized dataset with numerical ranges and masked sensitive attributes.
- Returns list of anonymized DataFrames.

anonymize:

- Reads dataset from CSV file.
- Takes user input for K and L.
- Calls partition_dataset and get_anonymize_dataset.
- Writes anonymized datasets to text files.

Main Flow:

Input Parameters:

- filename: CSV file with the dataset.
- User inputs for K and L values.

Reading Data:

- Reads dataset from CSV using pd.read_csv.

Partitioning:

- Calls partition_dataset for K-anonymity and L-diversity.
- Handles categorical and numerical columns.

Anonymization:

- Calls get_anonymize_dataset for anonymized datasets.
- Replaces numerical values, masks sensitive attributes.

Output:

- Writes anonymized datasets to text files (k_anonymity.txt, k_drawback.txt, median_problem.txt).
- k_anonymity.txt: Anonymized dataset.
- k_drawback.txt: Groups with same sensitive attribute values (for L=0 or 1).
- median_problem.txt: Groups with size >= 2K.

Runtime Calculation:

- Calculates and prints runtime.

Notes:

- Supports K-anonymity and L-diversity.
- Sensitive Column,categorical columns,numerical columns should be mentioned in code according your dataset.
- Handles both categorical and numerical columns.
- Tracks merged categories in merged_categories.
- Anonymized datasets written to separate files.

Important:

- Success of K-anonymity depends on dataset and parameter choices (K, L).