# InfoSum: Set-Intersections

## Background

InfoSum's platform is built to generate insights from any number of data sources without pooling the data in one location or revealing any identifiable information. To ensure that privacy is maintained the data stored in our platform will only ever contain anonymised data.

Information is imported into our platform as **datasets**, these **datasets** are lists of information about individuals which have been processed to ensure that personal identifying information (PII) has been removed. A **dataset** has multiple rows of data, with multiple columns. Each row of data has one or more **keys** which can be used to uniquely identify an individual within the **dataset**.

One type of **key** that can be used to identify individuals is **UDPRN. UDPRN** stands for Unique Delivery Point Reference Number which is a Royal Mail identifier for every unique delivery address in the UK.

## The Task

Write a program that will allow the comparison of two **datasets**. These **datasets** are provided in the form of a CSV file. These simple example CSV files contain a list of **UDPRN keys**.

The program should perform the following:

- Allow specification of the files to process
- Calculate and display the following:
    - The count of the **keys** in each file
    - The count of the distinct **keys** in each file
    - The count of the overlap of distinct **keys** between the two files (distinct overlap)
    - The product of the overlap of all keys between the files (overlap product)

Overlap Product is defined as the sum of the products of the overlapping keys, e.g.

```
Dataset 1: A B C D D E F F
Dataset 2: A C C D F F F X Y
Distinct Overlap = A C D F = 4
Overlap Product = A C C D D F F F F F F = 11
```

## Additional Considerations

- The program should be able to handle different files in csv format
- Consider how the program would work with larger files both in terms of number of rows/columns of data and number of unique key values
- If approximations are used, ensure the accuracy of the values is appropriately represented
- How you would validate the above requirements are met
- Feel free to ask any questions about this task

## Deliverables

1. Source code for the solution
2. Instructions on how to build and run
3. Appropriate documentation