**How to Read Duplicates Files**

In the process of matching the HMIS and Connecting Point data, I found probable duplicates within both data sets. I didn’t remove any of the duplicates, but instead created a file to identify them. Below is an explanation of the fields in the two datasets…

*Warning: Data must be kept in current sort order in order to keep duplicate groups together.*

**File: hmis\_client\_duplicates.csv (HMIS Data)**

|  |  |
| --- | --- |
| **Field** | **Description** |
| Subject Unit Identifier | Unique client ID for this dataset |
| Duplicates | if 0, then no duplicate client found  If 1, then first in a group of duplicates.  If 2+, then second in a group of duplicates and so on. |
| Duplicate ClientID | For rows with “Duplicates” value >=2, this field contains the unique ID of the first in the group of duplicates (i.e., the unique id for “Duplicates” =1). |

**File: cp\_client\_duplicates.csv (Connecting Point Data)**

|  |  |
| --- | --- |
| **Field** | **Description** |
| clientid | Unique client ID for this dataset |
| caseid | Caseid associated with each clientid |
| Duplicates | if 0, then no duplicate client found  If 1, then first in a group of duplicates.  If 2+, then second in a group of duplicates and so on. |
| Duplicate ClientID | For rows with “Duplicates” value >=2, this field contains the unique ID of the first in the group of duplicates (i.e., the unique id for “Duplicates” =1). |