Scripts to run in order:

1) “cleaning cls and mrip data 2016.R” – gets 2 tidy datasets, 1 for CLS and 1 for MRIP for 2016

-called cls\_tidy\_full.csv (all reports 2016-part of 2018) and mrip\_tidy\_16.csv

1. “cleaning cls and mrip data 2017.R” 1 tidy dataset for MRIP for 2017

- mrip\_tidy\_17.csv

3) “matching mrip and cls all possible matches 2016.R” - clean up and do a full join between cls and mrip data based on CLS ID and Wave/Year. – get 3 datasets cls\_tidy2\_full.csv (updated for full cls data), mrip\_tidy2\_16.csv (updated), and 2016\_all\_possible\_matches.csv

4) “matching mrip and cls all possible matches 2017.R” - clean up and do a full join between cls and mrip data based on CLS ID and Wave/Year. – get 2 datasets: mrip\_tidy2\_17.csv (updated), and 2017\_all\_possible\_matches.csv

5) “cleaning up all possible matches 2016” – cleans up all possible matches from 2016, outputs tidy\_all\_matches\_docks\_2016.csv which can be considered a dataset of the non-matches from 2016 since there are so few matches relative to all possible matches

6) “cleaning up all possible matches 2017” – cleans up all possible matches from 2017, outputs tidy\_all\_matches\_docks\_2017.csv which can be considered a dataset of the non-matches from 2017 since there are so few matches relative to all possible matches

7) “merging 2016 and 2017 all possible matches.R” – binds the rows of tidy\_all\_matches\_docks\_2016.csv and tidy\_all\_matches\_docks\_2017.csv together to get a large dataset of all possible matches from 2016 and 2017. – output all\_possible\_matches\_16\_17.csv

---- “source\_all\_possible\_matches.R” does steps 1-7 in one script, scripts 1-7 in *data\_processing\_for\_record\_linkage* folder

8) “distribution of variables from all possible matches 2017 and 2016.R” – get the distributions of variables from the non-matches from the combined all possible matches from 2016 and 2017 – not needed in flow at this point (3-20)

10) “distributions of linking variables from matches 2016 and 2017.R” – get the distributions of linking variables from the matches from the combined all possible matches from 2016 and 2017

11) “record\_linkage\_algorithm\_updated.R” – assigns a score to each possible match based on a record linkage algorithm (also implemented in this code), outputs tidy\_all\_matches\_docks\_rl.csv which is a dataset of all possible matches, scores from every linking variable for each record, and the summed score of each linking variable for all possible matches.

12) “obtain matches using record linkage.R” – gets at most one cls report per mrip intercept sample, writes these data to tidy\_all\_matches\_rl\_one\_per.csv, and each row has a record linkage score attached to it