Conversion and Data Wrangling Process

STEP 1: Extraction ⇢ **STEP 2:** Files Conversion & Concatenation in R ⇢ **STEP 3:** Randomly sample 20% of available data ⇢ **STEP 4:** Output

| Extract 2015 AIH .dbc files (100+ files) from public website. | → | Convert teach file into an R dataframe | → | Concatenate all dataframes and output as CSV | → | Load CSV to pandas and take a proportional random sample |

| Extract 2016 AIH .dbc files (100+ files) from public website. | → | Convert teach file into an R dataframe | → | Concatenate all dataframes and output as CSV | → | Load CSV to pandas and take a proportional random sample |

| Extract 2017 AIH .dbc files (100+ files) from public website. | → | Convert teach file into an R dataframe | → | Concatenate all dataframes and output as CSV | → | Load CSV to pandas and take a proportional random sample |

| Extract 2018 AIH .dbc files (100+ files) from public website. | → | Convert teach file into an R dataframe | → | Concatenate all dataframes and output as CSV | → | Load CSV to pandas and take a proportional random sample |

Concantrenate each year's random sample into one dataframe.

↓

Drop all empty or nearly empty columns. Forty-three columns are dropped, 69 remain.

↓

Save resulting dataframe as CSV for further analysis and exploration.