**DATA**

The data we will use comes from the EPA’s Enforcement and Compliance History Online (ECHO) system, which contains information from the Resource Conservation and Recovery Act Information System (RCRAInfo). This data is available at <https://echo.epa.gov/tools/data-downloads/rcrainfo-download-summary>. RCRAInfo tracks hazardous waste handlers, their activities, inspections of their activities, and enforcement and compliance actions taken. The data comes stored in six primary tables:

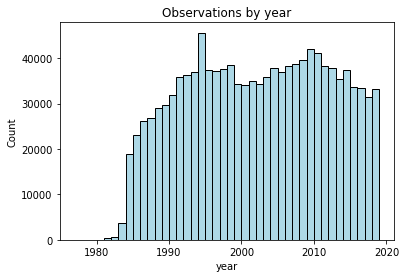
* RCRA\_FACILITIES – basic information about each RCRA site/hazardous waste handler, including name, location, and dummy variables that indicate what types of activities the facility engages in, its federal regulatory status, whether it transports hazardous waste, and whether it is an active site
* RCRA\_ENFORCEMENTS – count and description of enforcement actions taken against a given facility, the enforcement agency and date action taken, and dollar amounts of any settlements or punishments required
* RCRA\_EVALUATIONS – count and description of evaluations performed at a given facility, the evaluating agency and start date, and whether the evaluation resulted in a violation being found
* RCRA\_VIOLATIONS – detailed violation information, including the facility at which the violation occurred, its type, investigating agency, determination date, and the scheduled/actual date at which the facility brought the violation into compliance
* RCRA\_NAICS – the industrial activity classification/description of the facility, as defined by the Census Bureau (see below)
* RCRA\_VIOLATION/SNC HISTORY – dummy variables indicating whether the facility has unresolved violations and whether it has been designated as an unresolved significant noncomplier.

All tables contain a unique *ID\_NUMBER* field that can be used to link all datasets. This ID indexes observations at the facility/site level, meaning that it uniquely identifies observations in the RCRA\_FACILITIES dataset but may appear multiple times in other datasets (for example, *ID\_NUMBER* may appear multiple times in RCRA\_VIOLATIONS if the same facility commits multiple violations).

The main ancillary data source we will need is the Census Bureau’s North American Industry Classification System (NAICS), which is a standard industrial coding system that defines a facility’s activity type. This data is available at <https://www.census.gov/eos/www/naics/>. We will use this data in conjunction with the *NAICS\_CODE* field in the RCRA\_NAICS dataset to extract a description of each facility’s activity type, which may be a useful predictor in the models we will build.

In thinking about potential features, there are some fields (like count of previous violations, indicators of unresolved violations and noncomplier status) that could potentially be strong indicators of whether an evaluation would result in a violation. It could also be informative to investigate the facility’s status as a waste generator, hazardous waste transporter, and class of industrial activity.

A cursory review of the dataset provides an encouraging result: there appears to be a relatively consistent amount of data on a yearly basis over the span of the dataset. A simple merge of the RCRA\_EVALUATIONS, RCRA\_FACILITIES, and RCRA\_NAICS tables results in a dataset of 1,225,500 observations, which are reasonably evenly distributed over the years 1984-2018 (with only a few outliers that fall outside this range). Furthermore, there are very few null values of key variables. The only consistently null fields are *PMP\_AMOUNT, FMP\_AMOUNT, FSC\_AMOUNT,* and *SCR\_AMOUNT*, all found in the ENFORCEMENTS table. These fields represent the dollar amount of various fines and punishments levied on violating facilities, and it makes sense that there could be many null values as there are many possible enforcement actions beyond financial penalties.



enforcements

There are 332540 observations of 11 variables:

colname type pct\_null

0 ID\_NUMBER <class 'str'> 0.000000

1 ACTIVITY\_LOCATION <class 'str'> 0.000000

2 ENFORCEMENT\_IDENTIFIER <class 'str'> 0.000000

3 ENFORCEMENT\_TYPE <class 'str'> 0.000000

4 ENFORCEMENT\_DESC <class 'str'> 0.000069

5 ENFORCEMENT\_AGENCY <class 'str'> 0.000000

6 ENFORCEMENT\_ACTION\_DATE <class 'str'> 0.000000

7 PMP\_AMOUNT <class 'numpy.float64'> 0.955380

8 FMP\_AMOUNT <class 'numpy.float64'> 0.919856

9 FSC\_AMOUNT <class 'numpy.float64'> 0.997312

10 SCR\_AMOUNT <class 'numpy.float64'> 0.996364

Summary of numeric variables

colname mean median min max std\_dev count

0 PMP\_AMOUNT 71880.810855 10000.0 NaN NaN 768122.800148 332540

1 FMP\_AMOUNT 49024.297915 7500.0 NaN NaN 853275.571100 332540

2 FSC\_AMOUNT 101109.412472 12198.5 NaN NaN 754339.029781 332540

3 SCR\_AMOUNT 34949.278031 9000.0 NaN NaN 113164.781465 332540

-----------------------------------------------------------------------

evaluations

There are 991196 observations of 8 variables:

colname type pct\_null

0 ID\_NUMBER <class 'str'> 0.0

1 ACTIVITY\_LOCATION <class 'str'> 0.0

2 EVALUATION\_IDENTIFIER <class 'str'> 0.0

3 EVALUATION\_TYPE <class 'str'> 0.0

4 EVALUATION\_DESC <class 'str'> 0.0

5 EVALUATION\_AGENCY <class 'str'> 0.0

6 EVALUATION\_START\_DATE <class 'str'> 0.0

7 FOUND\_VIOLATION <class 'str'> 0.0

Summary of numeric variables

None

-----------------------------------------------------------------------

facilities

There are 1041254 observations of 15 variables:

colname type pct\_null

0 ID\_NUMBER <class 'str'> 0.000000e+00

1 FACILITY\_NAME <class 'str'> 5.762283e-06

2 ACTIVITY\_LOCATION <class 'str'> 0.000000e+00

3 FULL\_ENFORCEMENT <class 'str'> 0.000000e+00

4 HREPORT\_UNIVERSE\_RECORD <class 'str'> 9.315691e-05

5 STREET\_ADDRESS <class 'str'> 1.306117e-04

6 CITY\_NAME <class 'str'> 9.795881e-05

7 STATE\_CODE <class 'str'> 9.603805e-07

8 ZIP\_CODE <class 'str'> 1.930365e-04

9 LATITUDE83 <class 'numpy.float64'> 5.653500e-01

10 LONGITUDE83 <class 'numpy.float64'> 5.653500e-01

11 FED\_WASTE\_GENERATOR <class 'str'> 8.250629e-03

12 TRANSPORTER <class 'str'> 2.823519e-04

13 ACTIVE\_SITE <class 'str'> 0.000000e+00

14 OPERATING\_TSDF <class 'str'> 0.000000e+00

Summary of numeric variables

None

-----------------------------------------------------------------------

naics

There are 426042 observations of 3 variables:

colname type pct\_null

0 ID\_NUMBER <class 'str'> 0.0

1 ACTIVITY\_LOCATION <class 'str'> 0.0

2 NAICS\_CODE <class 'str'> 0.0

Summary of numeric variables

None

-----------------------------------------------------------------------

violations

There are 594380 observations of 8 variables:

colname type pct\_null

0 ID\_NUMBER <class 'str'> 0.000000

1 ACTIVITY\_LOCATION <class 'str'> 0.000000

2 VIOLATION\_TYPE <class 'str'> 0.000000

3 VIOLATION\_TYPE\_DESC <class 'str'> 0.000000

4 VIOL\_DETERMINED\_BY\_AGENCY <class 'str'> 0.000000

5 DATE\_VIOLATION\_DETERMINED <class 'str'> 0.000000

6 ACTUAL\_RTC\_DATE <class 'str'> 0.021666

7 SCHEDULED\_COMPLIANCE\_DATE <class 'str'> 0.000000

Summary of numeric variables

None

-----------------------------------------------------------------------

viosnc\_history

There are 2318155 observations of 5 variables:

colname type pct\_null

0 ID\_NUMBER <class 'str'> 0.0

1 ACTIVITY\_LOCATION <class 'str'> 0.0

2 YRMONTH <class 'numpy.int64'> 0.0

3 VIO\_FLAG <class 'str'> 0.0

4 SNC\_FLAG <class 'str'> 0.0

Summary of numeric variables

None

-----------------------------------------------------------------------