# Appendix B: Data Validation

Since 2010, California has produced more than 400,000 manifests each year (Appendix Figure 1). With this number of documents, errors could be anticipated. Analysis of weights per container types did reveal errors that cause large outliers in the Hazardous Waste Tracking System (HWTS). To account for this, DTSC used a statistical approach to get an understanding of the data within HTWS, including the number of statistical errors within HWTS.
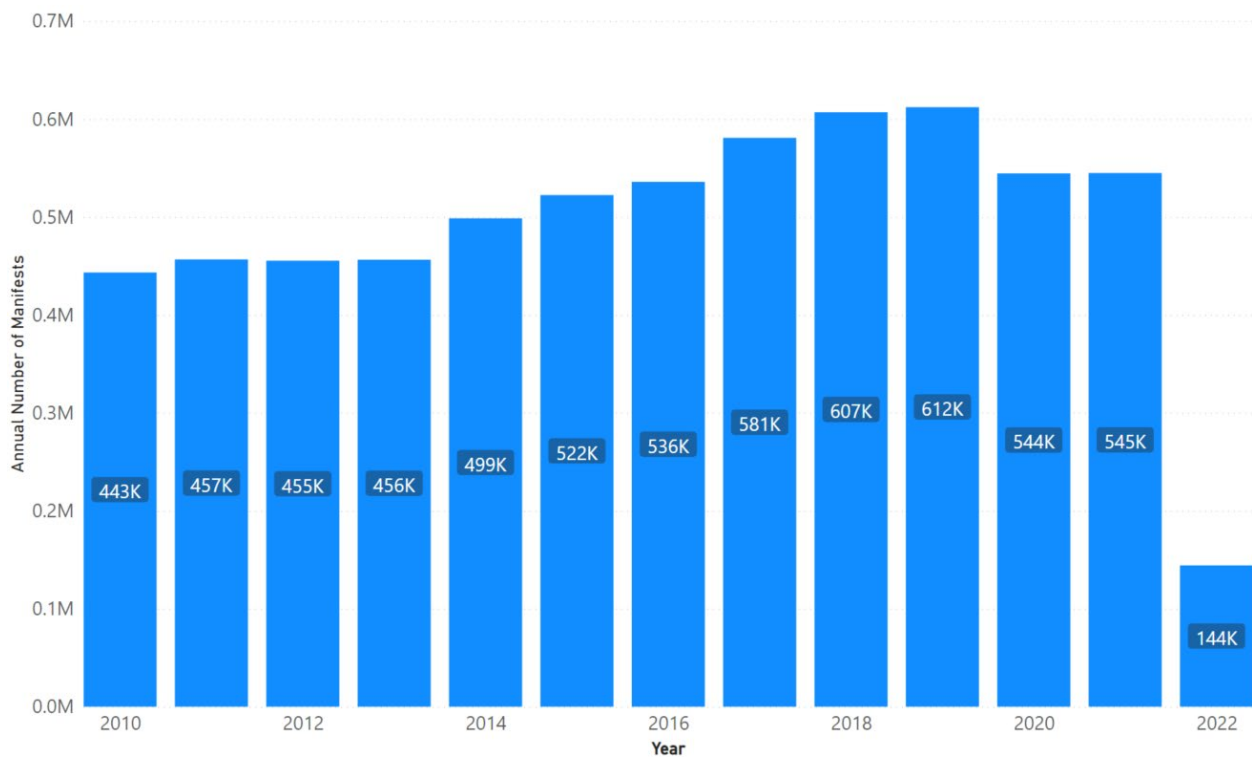
Appendix Figure 1: Annual Number of Manifests in California

To get a better understanding of the HWTS data, DTSC analyzed each of the 13 shipping container types used to ship hazardous waste:

- Burlap, cloth, paper, or plastic bags
- Fiber or plastic boxes, cartons, cases
- Metal boxes, cartons, cases (including roll-offs)
- Wooden boxes, cartons, cases
- Cylinders
- Fiberboard or plastic drums, barrels, kegs

- Metal drums, barrels, kegs
- Dump trucks
- Wooden drums, barrels, kegs
- Hopper or gondola cars
- Tank cars
- Portable tanks
- Cargo tanks (tank trucks)

When reviewing this data within HWTS, DTSC found some manifests with errors that resulted in extreme outliers among the reported weights of hazardous waste when compared with other shipments in the same containers. The most extreme example was a dump truck shipment that reported to have shipped more than 76,000 tons of hazardous soil (Manifest: 010275196FLE). This is a major deviation from the median dump truck shipment weight of 25.4 tons. Upon inspection, DTSC discovered that poor penmanship led to an incorrect amount of waste soil being entered on the manifest.

Consequently, to ensure an accurate accounting of the amount of hazardous waste being generated in California, DTSC developed a method to detect and correct data outliers. Once the raw data was extracted from HWTS, every container type's median, mean, and standard deviation was determined in tons. This threshold was the container type's median weight plus three times the standard deviation. If the manifested weight was under this threshold, the original weight on the manifest was retained. If the weight was greater than this threshold, each container's weight was adjusted to the median weight of that container type. The only exception to this was for cylinders (container code: CY), where the median value was 0. To account for this, the mean value was used.

Here are the 13 shipping container types and the values used for data validation, in tons:

- Burlap, cloth, paper, or plastic bags (Median: 0.20, SD: 11.25)
- Fiber or plastic boxes, cartons, cases (Median: 0.07, SD: 108.97)
- Metal boxes, cartons, cases (including roll-offs) (Median: 16.65, SD: 212.74)
- Wooden boxes, cartons, cases (Median: 0.33, SD: 14.64)
- Cylinders (Mean: 0.49, SD: 2.90)
- Fiberboard or plastic drums, barrels, kegs (Median: 0.03, SD: 1.25)
- Metal drums, barrels, kegs (Median: 0.11, SD: 7.07)
- Dump truck (Median: 25.38, SD: 418.75)
- Wooden drums, barrels, kegs (Median: 0.13, SD: 0.05)
- Hopper or gondola cars (Median: 84.60, SD: 1,543.20)
- Tank cars (Median: 79.82, SD: 38.76)

- Portable tanks (Median: 1.08, SD: 12.94)
- Cargo tanks (tank trucks) (Median: 7.81, SD: 64.41)
- Uncategorized / Error (Mean: 5.30, SD: 140.61)

After this data analysis was completed, of the 6.40 million manifests produced since 2010, 2184 (0.034%) were found to have likely outliers and received a data adjustment, per the method described above. With these 2184 manifest corrections, the resulting decline in manifested hazardous waste weight was 9.25 million tons since 2010

It was found that in some years there were hundreds of manifests with potential errors while other years had only tens of manifests with potential errors. Appendix Figure 2 shows the count of statistical errors per year.
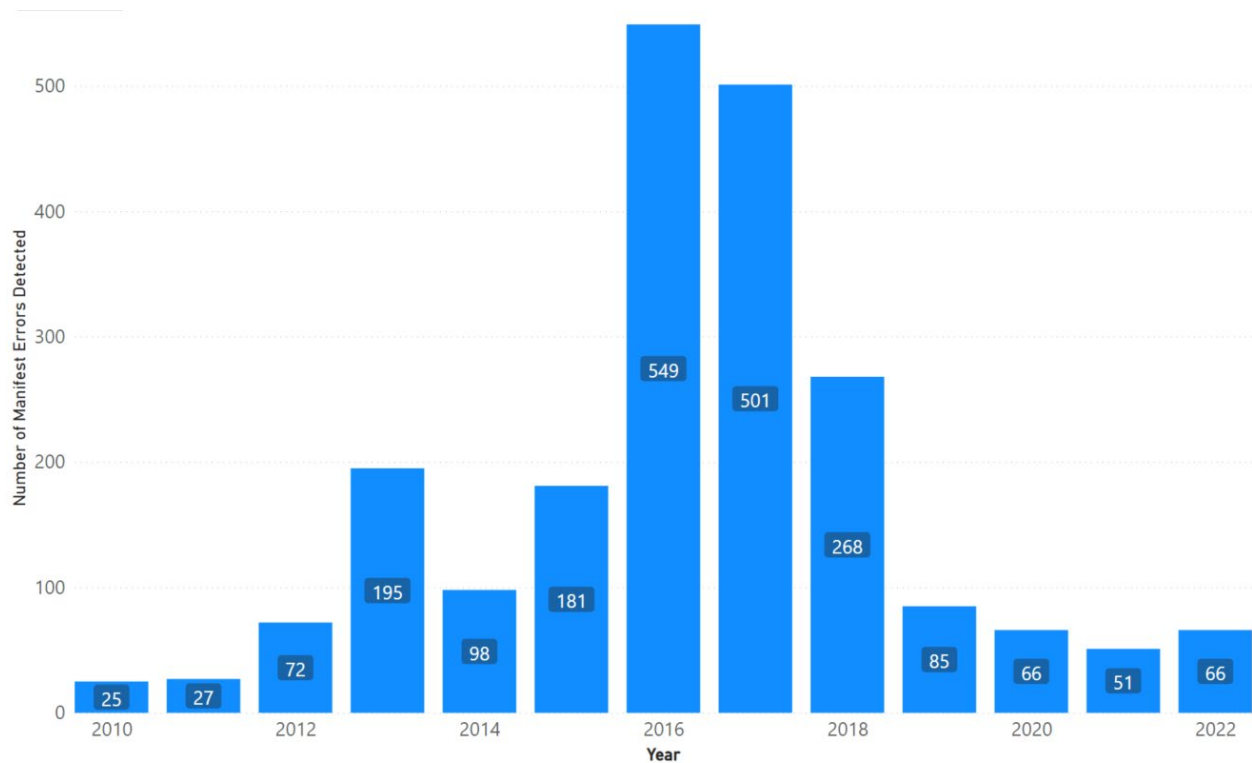
Appendix Figure 2: Annual Number of Manifests with Statistical Outliers