**Data Extraction**

Dataset : 60000 rows , 12 columns

* Preprocessing steps:
  + Renaming columns
  + Convert dtypes (?)
* Defining debug and working datasets 10000 for debug, 60000 for working
  + [Shenglan Li](mailto:sl182@illinois.edu) Take the original dataset (no shuffling) as working dataset

**Data Exploration**

* The rename columns and convert dtypes have been done before **delete and replace by working\_dataset**
* Looking at the NA values in the dataset: they are found in Data\_value, Footnote, Footnote\_symbol and Predicted\_value;
* Looking at the unique values of Indicator: the 'indicator' includes not just drug type, but also 'Number of Drug Overdose Deaths', 'Number of Deaths' and 'Percent with drugs specified'. The 'Number of Drug Overdose Deaths' is the feature we want to predict. Also, the 'Natural & semi-synthetic opioids, incl. methadone (T40.2, T40.3)' and 'Natural, semi-synthetic, & synthetic opioids, incl. methadone (T40.2-T40.4)' overlap with other drug classes.
* There are rows based on State=US; we drop them because they sum the number of deaths in each state.
* Converting date types to Year and Months
* Check the correlation between different features by pairplot
* For df1, we observe the trend of yearly overdose death going up and build bar plot to check the difference of overdose death within each state, for unnormalized data, CA, FL, OH and PA has the highest number while for data normalized by population of each state, DC has the highest number.
* Construct df2 to check the trend and distribution of death number within each drug indicator. The Opioids has the highest death number.
* For df2, we also focused on looking at overall data and comparing the difference between real data and the predicted data. Predicted data is highly correlated with real reported data. The difference between real data and the predicted data is not very big. But basically, the difference is increasing over time. The fluctuation is quite intensive after 2022, which may because that provisional estimates of drug overdose deaths have traditionally been reported 6 months after the date of death, but starting in February 2022, the 6-month lag was shortened to 4 months.

**Base Learning**

Dataset:

* Keep features we need: Year, Data\_Value(OD count), Percent\_Pending\_Investigation, State(48 predictors), Month(12 predictors), Indicator (drug type)(7 predictors)
* Remove State LA, ND, NE, PA due to not having data for courses of drug overdose death; Remove Indicators which do not make sense (i.e., Number of Deaths, Number of Overdose Deaths)
* Df\_processed: Min max scaler for numerical variables and dummies for categorical variables(State, Month, Indicator).

Run linear regression:

* Learning rate: 0.01
* Test\_size: 0.2
* Check correlation between features: approximately no significant correlation within them.
* **Large correlation between predicted value and data value, which causes the prediction to be perfect.(Hell no!)**
* **Large errors for linear regression after removing the data value feature.**

**Deep Learning**

* **Load the correct, preprocessed dataset with the right columns (remove the cells mentioning of df2)**

**Feature Importance**