Summary of Data Cleaning

We cleaned the original dataset, i.e., `gtd.csv` and saved the cleaned dataset as an output csv file named `gtd_cleaned.csv`. Below are steps and approach we used in this notebook in concise manner:

Dropping columns

- Checked for duplicate values (if any) in uncleaned data, fortunately, we had none.
- Next, we checked for missing values in our data and got a total missing values for each column.
- Dropped all the columns with high volume of missing values (more than **50%** missing values). Started with **135** columns, after dropping columns, ended up with **77** columns.
- `weapdetail` column had too many unique values alongwith high volumne of missing data (**41%**), so we dropped it too.
- `scite1` cited the first source that was used to compile information on the specific incident, it was not very relevant for our analysis, so dropped it. Similar goes with `summary` column.
- Both `weapsubtype1` and `weapsubtype1_txt` were categorical variables and represented same information, we only kept only one of them, i.e., `weapsubtype1_txt` as it gives textual information of weapon subtype, `weapsubtype1` gave the numerical categorical info.
- Dropped `nperps`, `weapdetail`, `scite1`, `summary`, `weapsubtype1`, `targsubtype1`, `natlty1`, `weaptype1`, `eventid`, `country`, `region`, `attacktype1`, `targtype1`, `specificity` and `individual` column as they either were not very relevant or had other alternative columns present which imparted same information as them.

Missing Value Imputation

Next, we looked deeper into the remaining columns for missing value imputation and other modifications:

Numerical columns

- for numerical columns, imputed missing values of columns `multiple` with 1, `latitude`, `guncertain` and `longitude` with their respective modes (as percentage of missing values was not very high (around 2.5%).
- for columns such as `nperpcap`, `claimed`, `nkill`, `nkillus`, `nkillter`, `nwound`, `nwoundus`, `nwoundte` and `ishostkid`, imputed missing value with **-99** as it is being used for indicating missing value.

Categorical. columns

- for columns `city`, `targsubtype1_txt`, `corp1`, `target1`, `natlty1_txt` and `weapsubtype1_txt`, imputed missing values with **Missing Info** to avoid any confusion with non-null values.

Datatype Conversion and Mapping

- After this we changed the datatype of some float columns which were supposed to integer datatype but were misclassfied into the wrong datatype, these columns included `nperpcap`, `nkill`, `nkillus`, `nkillter`, `nwound`, `nwoundus` and `nwoundte`
- For some columns which were having integer values (0, 1 and -99), we mapped them as {1:"Yes", 0:"No", -99:"Unknown", -9:"Unknown"} as per the convention given in `Codebook.pdf`

Renaming the columns

- Finally, we used the `Codebook.pdf` to rename the columns to more suitable and interpretable names.
- Saved and generated the cleaned data into `gtd_cleaned.csv`