**Data Preprocessing**

1. Removing redundant columns
   1. Location
      1. Can be removed as it serves as a combination of the Latitude and Longitude
   2. Year
      1. 2021 is the only year listed, therefore can be removed
   3. Case Number
      1. Arbitrary tracking key value, not indicative feature of the dataset
   4. ID
      1. Same as Case Number
   5. Block
      1. Information might be too correlated to features like beat, district
      2. X and Y coordinates/Longitude and Latitude are more effective and require less data preprocessing to be viable
   6. Updated On
      1. Not a value indicative of the crime therefore not valuable to assessment of the crime
2. Columns Included
   1. **IUCR -> Illinois Uniform Crime Reporting code**
      1. **Approx. 410 unique IUCR**
      2. **Feature Column includes letters and numbers, requires transformation into quantitative data**
      3. **Consider summing these columns into simpler categories, which are highly similar to that of the Primary Type category, such as below:**
      4. **https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e**Chart, pie chart

         Description automatically generated
   2. **Primary Type**
      1. **Summarized qualitative data of the IUCR report.**
      2. **Data contains ~31 different Primary Types**
      3. Chart, histogram

         Description automatically generated**Requires quantifying**
3. **Location Description**
   1. **125 distinct quantitative descriptions of where incident occurred**
   2. **Requires quantifying**
4. **Arrest**
   1. **Boolean value indicating whether an arrest was made or not**
   2. **Set arrest occurred to 1 and not occurred to 0**
5. **Domestic** 
   1. **Indicates whether the incident was domestic-related as defined by the Illinois domestic violence act using**
   2. **Set domestic-related occurrences (True) to 1 and otherwise (false) to 0**
6. **Beat/District/Ward/Community Area/X and Y/Latitude and Longitude**
   1. **Beat -> smallest geographic unit of police distributive breakdown**
   2. **District -> larger geographic breakdown of police (22 districts in Chicago)**
   3. **Ward -> area broken up by city council districts**
      1. **Map:** [**https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Wards-2015-/sp34-6z76**](https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Wards-2015-/sp34-6z76)
   4. **Community Area -> designated “community areas” of Chicago, broken up into a total of 77 community areas**
      1. **Community Area boundaries are not political (not susceptible to redistricting or red-lining/gerrymandering), so very consistent for long term collection and analysis.**
      2. [**https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6**](https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6)
   5. **X and Y -> X and Y coordinates of the location where the incident occurred as mapped to the State Plane Illinois East NAD 1983 projection (basically an isolated Latitude Longitude for the City of Chicago).**
      1. **Partial reprojection of actual location occurs to redact exact location of incident. Reprojection done to the center of the nearest block.**
   6. **Latitude and Longitude -> Global latitude and longitude of incident location**
      1. **Partial reprojection of the actual location occurs to redact exact location of incident. Reprojection done to the center of the nearest block.**
   7. **FBI Codes**
      1. **Crime classification as outlined in the FBI’s National Incident-Based Reporting System.**
      2. **Text-data format**
      3. **Contains 26 unique FBI codes, similar in meaning to that of the Primary Type column**

**Recommended Column Usages**

* **Primary Type**
  + **Mapped randomly to an arbitrary range of quantifiable values based on their distinct counts**
  + **Utilize a translation table to transform data**
* **Location Description**
  + **Mapped randomly to an arbitrary range of quantifiable values based on their distinct counts**
  + **Utilize a translation table to transform data**
* **Domestic**
  + **Boolean value true or false mapped to 1 (True) or 0 (False)**
* **Arrest**
  + **Boolean value true or false mapped to 1 (True) or 0 (False)**
* **Location Column**
  + **Use Community Area identifier**
    - **No data transformation necessary, each community area corresponds to a number between 1 and 77**