**Assignment 2 Part 1**

**INSIGHTS After Profiling LA Crime Dataset**

**Issues With Data**

* Embedded newlines in some cells
* Column names to be renamed properly
* >256 Characters
* Missing Values (more than 10% of values are missing for some fields)
* CSV is comma Delimited and some values have commas, so while loading the file unwanted columns are created

**Cleaning Strategy**

* Remove newline character by “”
* Used Dynamic Rename module to rename column names
* Trim values
* Create a CSV file using Alteryx and delimited character be something unique like ‘;’
* Handling Missing Values:
  + Null Values can create inconsistencies when the significance of the column is higher
  + Column like Mocodes , some descriptions columns can be null
  + Columns like Age, Gender and Race can be handled in talend using vmap and replacing the null value with 0 or ‘X’