**Data Profiler Logic and UI/UX**

**Logic**

* Data Profiling
  + Purpose: Scan data sources and compile statistics based on the data
  + Create New Profile
    - Profile Name
    - Add imported data source
    - ~~add email addresses (multiple) for finished profile email (will email the report)~~
      * allow user to export via native connectors the profile summary
      * allow user to export profiled data source
    - Description
  + There will be 5 tabs based on the dimensions of Data Quality. We will calculate a score that will be displayed on the Tab (Example: Accuracy - 82%), also highlighted in yellow and red when scores are low. In each tab, highlight sections where data quality issues exist. Also, an overall Data Quality score should be calculated
  + Tab 1 Accuracy
    - Pattern detection
      * This will apply a list of RegEx to the columns and report the qty of data that is valid (pattern detected) or not valid (not detected)
    - MAX LEN for a column
    - Character detection
      * Contains Numbers, numbers only, Contains Letters, Letters only, Contains letters and numbers
      * Punctuation
        + I would like to report record counts for each punctuation character where the count is bigger than 0
      * Leading spaces
      * trailing spaces
      * non-printable characters
      * Outlier detection
        + Example: a column contains 2 distinct values over 1000 records, however, 1 additional value exists for a few records. These are outliers
      * Null vs Filled
  + ~~Tab 2 Completeness~~
  + Tab 3 Uniqueness
    - Distinct (how many distinct values are in a column)
    - Geo Map (this is low priority), if we can map ZIP/Postal codes to USA and Canada to show visualization of map with data points
  + Tab 4 Conformity
    - Data Type (String, integer, data, double, etc)
    - Date Format
  + Tab 5 Precision
    - Numerical statistics if a column is numeric (Min, Max, Mean, Median, Mode and Extreme)
  + Data preview is important, when a customer selects a metric, the data should filter and show only records affected by case
    - For Distinct records option, we should show pivot table with the distinct value, count of repeat and % of total data source this count occupies
    - user should be able to export data using the native data source connectors
  + Histogram, allow user to click to filter based on histogram bars
  + Need a nice report in PDF or Excel

**UI**

* The idea behind UI is to have the following
  + Tabs, views or sections: Accuracy, Completeness, Uniqueness, Conformity and Precision
  + After Profiler is run, each tab will have the name and the score that I will help you calculate
  + There will be a top section with static info that will not change even if you change the tab view: Total records, Overall data source Score,
  + Allow the user to export profile
  + When a user clicks on a statistical value, we should show the records that are affected.
    - We can show in a window on screen or
    - we can show a popup windows
    - in all cases, the user should be able to export the records
  + Display
    - Accuracy
      * Data Frame for various
        + Column headers
        + Pattern Detection

Place a button somewhere that says something like “Show Pattern Graph”. Make this a bar graph with 3 bars, total records, valid and invalid

Mouse over logic to show values

Click on bar shows records

Also, there should be a drop down menu to show other detected patterns

order by count of detected/valid fields

* + - * + MAX LEN
        + Leading spaces
        + trailing spaces
        + non-printable characters

Button for bar graph for Leading, trailing and non-print

* + - * + Null
        + Filled

button to display a pie graph nulls vs filled with mouse over logic

* + - * + Outlier Detection

Lets incorporate DBscan and see what happens.

Report the count of Noise Points.

https://towardsdatascience.com/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623#:~:text=It%20is%20the%20difference%20between,IQR)%20or%20boxplot%20upper%20whisker.

* + - * Data Frame for punctuation
        + only show punctuation that has detections
        + order by count of detection from most to least
        + Column 1: character
        + Column 2: count
        + column 3: % of records affected

sum(affected records) / total records

* + - * Data Frame
      * Data Frame for data source records and filter logic
    - Uniqueness
      * Distinctness
        + Lets use a histogram for this to show the most repetitive data
        + Also a DF that will show the distinct and their counts. Screenshot below
      * Geo map based on Zip codes for US (all states including PR) and canada. I will check if other countries use zip codes.
    - Conformity
      * data frame with
        + Column Data Type (String, integer, data, double, etc)

4 extra columns:

Valid if same format

Valid percentage of column

invalid if not same

Invalid Percentage of column

* + - * + Date Format (Only if date column, N/A if not datetime column)

Show the date format

Valid if same format

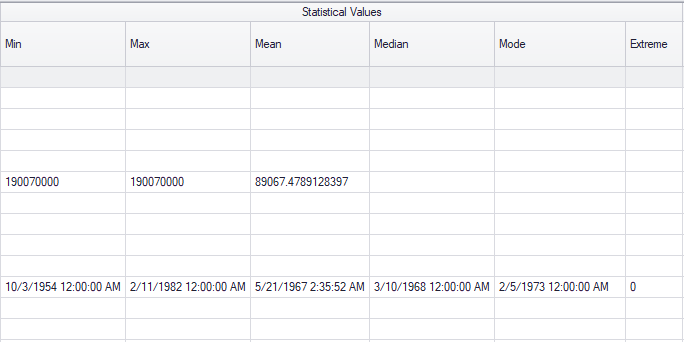
Valid percentage of column

invalid if not same

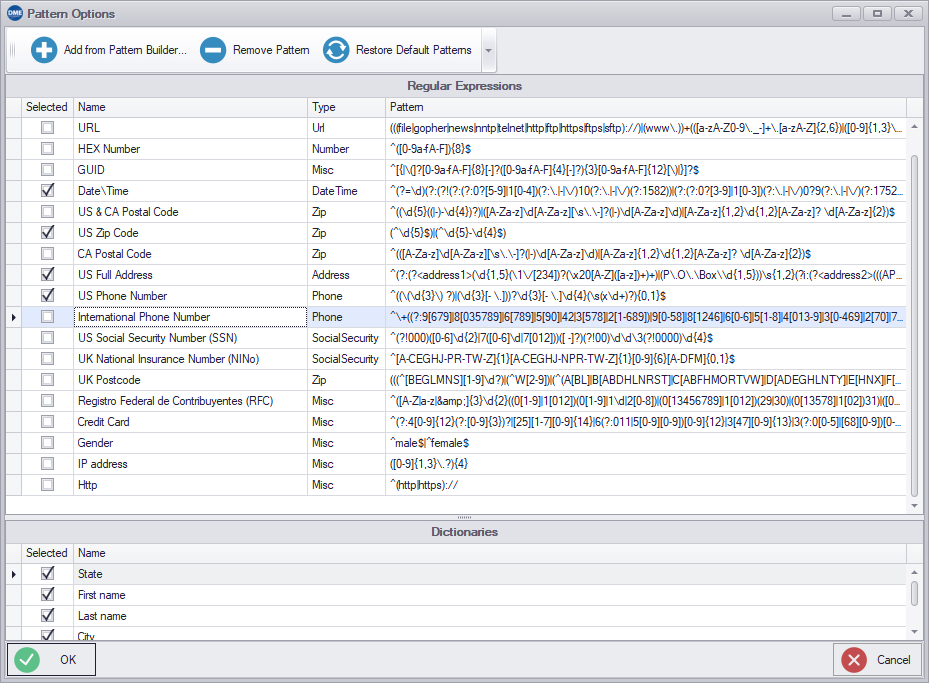
Invalid Percentage of column

* Overall Data Source Score
  + TR = Total Records
    - Accuracy: total 50%
      * Nulls: Total 20%
        + If 40% of TR are blank or null then 0%, else 20%
      * Leading or Trailing spaces or non-printable chars: total 10%
        + If 20% of TR have leading or trailing spaces or non-printable chars then then 0%, else 10%
      * Outliers: Total 10%
        + If TR has more than 20% outliers, then 0% else 10%
      * Char detection: Total 10%
        + If TR contain more than 20% of ,/\*%:;#!)^`~\|[{]}=(+¨?<> then 0%, else 10%
      * Pattern Detection Total 10%
        + if best pattern for any column has more than 30% invalid values in that column, then 0, else 10%
    - Uniqueness: Total 30%
      * TR over 40% duplicate, then 0% else 30%
    - Conformity: Total 20%
      * if data type or date type has more than 10% invalid values, then 0%, else 20%

**Precision UI example**

****

Sample Regex Management for profiler



distinct example

