OVERVIEW

The Credit Card CVV Plugin (CCCP) takes in a column name and dataset column values that are assumed to be converted to a list of strings. The CCCP first removes any leading and trailing spaces in the strings. The CCCP then removes any null (None) values or any string values denoting a null value ('NA', etc.) found in the list. Then, the CCCP returns a score for each value based on how well the CCCP is confident that that value is a credit card CVV number. After scores have been assigned, any outlier scores (anything not within 2 standard deviations from the mean) are removed before taking and returning the final average of the scores (the final confidence score for the entire column).

FUNCTION DEFINITIONS

DEPENDENCIES: re

get_confidence_score()

The *get_confidence_score()* function takes in a *string* called *col_name* and a list of *strings* called *col_vals*. *col_vals* is presumed to be a dataset column that has been converted to a list of *strings*. *col_name* is presumed to be the dataset column name. The function first checks to see if '*cvv', 'credit', or 'card'* is detected within the column name. If so, the *col_name_check* is set to *True* and *False* otherwise. The function then creates an empty *double* list called *scores* to contain the confidence score for each value in *col_vals*. The function then modifies *col_vals* by calling on the *remove_lead_trail_space()* then the *remove_null()* function. After, the function iterates through each value in *col_vals* and calls on the *get_elem_score()* function to append each score returned from that function to the *scores* list. Afterwards, the *remove_outliers()* function is called on the *scores* list. Finally, the average (which is *double* type) of the *scores* list is returned (this is the final confidence score for the entire column).

get_elem_score()

The *get_elem_score()* function takes in a bool called *col_name_check* and a *string* called *elem*. *col_name_check* is presumed to be *True* if '*cvv', 'credit', or 'card'* is detected in the column name and *False* otherwise. *elem* is presumed to be a value from a list of *strings*. If *col_name_check* is *True*, then a score multiplier of x1.0 is set. If not, the multiplier is set to x0.5. The function then uses regular expressions in each if-statement to determine what confidence score (which is *double* type) to return. *See the REGEX section for more information on the confidence score.*

For whatever non-zero confidence score is chosen, a score multiplier will be applied. This is set to *x1.0* if the column name was detected to have '*cvv', 'credit', or 'card'* in it and set to *x0.5* if not.

remove_null()

The *remove_null()* function takes in a list of *strings* called *col_vals*. *col_vals* is presumed to be a dataset column that has been converted to a list of *strings*. The function uses list comprehension to return a list where any *None* values have been removed and where any strings denoting null values have been removed. The list of strings that denote values are noted as: *['NA', 'N/A', 'na', 'n/a', 'Na', 'N/a']*.

remove_lead_trail_space()

The *remove_lead_trail_space()* function takes in a list of *strings* called *col_vals*. *col_vals* is presumed to be a dataset column that has been converted to a list of *strings*. The function uses list comprehension to return a list where all the string elements have leading and trailing space removed.

remove_outliers()

The *remove_outliers()* function takes in a list of *doubles* called *scores*. *scores* is presumed to be the list of confidence scores of each value in a dataset column. The function first creates an empty *set* to contain outliers found. The function then calculates the average and the standard deviation of *scores*. Then the function iterates through each score in *scores* and determines whether the score is within 2 standard deviations from the average. If the score is not found within that range, that score is added to the outlier list. Because 95% of values in any distribution (*Statistic Found Here)* fall within that 2 standard deviation range, generally only extreme outliers will be removed. Finally, list comprehension is used to remove any found outliers from the *scores* list then that modified list is returned.

REGEX EXPLAINED

Regex Form	Confidence Score
\d{3}|\d{4}	100.0

Generally, there is one regex expression per confidence score category.
If none of the above formats are matched, a confidence score of 0.0 is returned.

This regular expression is used with the *re.fullmatch()* function therefore any additional or missing items from the regular expression will cause a score of 0.0.

CVV numbers can either be 3 or 4 digit numbers. Because the format is so general *(3 or 4 digit numbers can easily signify an ID, etc.)*, the column name is highly important to distinguish between a CVV number and a non-CVV number.

If *'cvv', 'credit', or 'card'* is found in the column name then a score multiplier of x1.0 is set. So if the regular expression is matched and the column name matches, then a score of 100.0 is returned. If those numbers are not found in the column name then a score multiplier of x0.5 is set. So if the regular expression is matched and the column name does not match, then a score of 50.0n is returned.

These score multipliers and confidence score pairings are subject to change either to team majority vote or mentor discretion.

IMPLICATIONS FOR FUTURE

Since credit card information is highly sensitive information, we would like to add a flag that denotes this specific data type (credit card CVV) as sensitive.

Since this plugin uses the same format as other plugins that use regular expression, we would like to have one .txt file containing all data types' regular expressions and one Python file that handles all regular expression cases instead of having separate Python files for each data type.