OVERVIEW

The Phone Number Plugin (PNP) takes in a dataset column that is assumed to be converted to a list of strings. The PNP first removes any null (None) values or any string values denoting a null value ('NA', etc.) found in the list. Then, the PNP returns a score for each value based on how well the PNP is confident that that value is a phone 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

---

getConfidenceScore()

The *getConfidenceScore()* function takes in a list of *strings* called *col*. *col* is presumed to be a dataset column that has been converted to a list of *strings*. The function then creates an empty *double* list called *scores* to contain the confidence score for each value in *col*. The function then modifies *col* by calling on the *removeNull()* function. After, the function iterates through each value in *col* and calls on the *getElemScore()* function to append each score returned from that function to the *scores* list. Afterwards, the *removeOutliers()* 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).

getElemScore()

The *getElemScore()* function takes in a *string* called *elem*. *elem* is presumed to be a value from a list of *strings*. The function then uses regular expressions in each if-statement to determine what confidence score (which is *double* type) to return. *See the PHONE NUMBER CONVERSIONS section for more information.*

removeNull()

The *removeNull()* function takes in a list of *strings* called *col*. *col* 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']*.

removeOutliers()

The *removeOutliers()* 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.

PHONE NUMBER CONVERSIONS

From the phone number formats in this article, https://blog.insycle.com/phone-number-formatting-crm, these phone number formats were chosen:

Phone Number Format	Regex Form	Confidence Score
(555)555-5555	^\(\d{3}\)\s?\d{3}-\d{4}$	100.00
(555) 555-5555	^\(\d{3}\)\s?\d{3}-\d{4}$	100.00
+555-555-5555	^\+1?\s?\d{3}[\.-]\d{3}[\.-]\d{4}$\|^\+1\d{10}$	80.00
+1 555.555.5555	^\+1?\s?\d{3}[\.-]\d{3}[\.-]\d{4}$\|^\+1\d{10}$	80.00
+15555555555	^\+1?\s?\d{3}[\.-]\d{3}[\.-]\d{4}$\|^\+1\d{10}$	80.00
555.555.5555	^(1-)?\d{3}[\.-]\d{3}[\.-]\d{4}$	60.00
1-555-555-5555	^(1-)?\d{3}[\.-]\d{3}[\.-]\d{4}$	60.00
555-555-5555	^(1-)?\d{3}[\.-]\d{3}[\.-]\d{4}$	60.00
555 555 5555	^\d{3}\s\d{3}\s\d{4}$	40.00
5555555555	^\d{10}$	20.00

Generally, there is one regex expression per confidence score category.

There is **not** the case where a phone number matches two or more confidence score categories.

If none of the above formats are matched, a confidence score of 0.00 is returned.

The confidence scores for each format was chosen due to the specificity of the format. For example, the *(555) 555-5555* format generally does not denote anything except for phone numbers (hence the *100.0* confidence scores) whereas a pure 10-digit number (*5555555555*) could denote a phone number but also could very well represent a generic ID (hence the *20.0* confidence score). *These confidence score pairings are subject to change either to team majority vote or mentor discretion.*

IMPLICATIONS FOR FUTURE

In the future, we would like to handle international phone numbers instead of just U.S. phone numbers. This means adding more regular expression formats or using an external API to check if an item is a valid phone number or not.