OVERVIEW

FUNCTION DEFINITIONS

---

get_confidence_score()

The *get_confidence_score()* 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 *remove_null()* function. After, the function iterates through each value in *col* 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 *string* called *elem*. *elem* is presumed to be a value from a list of *strings*. The function then uses regular expressions in an if-statement to determine what confidence score (which is *double* type) to return.

The regular expression checks to see if there is a single letter in the string. If there is a single letter, a confidence score of 100.0 is returned. Otherwise, a 0.0 is returned. There is no intermediate value.

remove_null()

The *remove_null()* 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.

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.

IMPLICATIONS FOR FUTURE