State Plugin Documentation

Overview

The State Plugin helps preprocess and analyze a dataset column to calculate a confidence score for how likely each value in the column is a valid US State. The plugin uses a column name and a database to evaluate the results.

Dependencies

Regular Expression module - **Re**
statistics module
state_city_database.py / us_states- a Python file that contains a dictionary us_states containing all the US states.

Function Definitions

get_confidence_score()

The get_confidence_score() function takes in the column name as a string and a sample of the dataset column as a list of strings.

The function first checks if the column name contains the substring "state" or "states". A boolean variable “col_check” acts as a checker that is set to True if the column name matches the condition.

The next step is to preprocess the data: the list of values is sent to functions that remove all the null values and leading/trailing spaces from the data.

The function then initializes a list of scores to keep track of all the confidence score values for the column. Each value is checked through the valid_state() function and the scores list is appended accordingly. Finally, a function removes any outliers in the results and returns a median value which is the final confidence score of the entire column.

Variables used:

col_name- column name (String)
col_values- column values (list of Strings)
col_check- column name checker (Boolean)
scores- confidence score for each value (list of Integers)
valid_state_check- valid state checker (Boolean)
Input- col_name, col_values
Output- median value of scores (float)

valid_state() /state_city_database.py

This function checks each value of the column through the state_city_database.
This database contains a dictionary of all the 50 US states as keys and its abbreviation as values. Example: {“Californa”:” CA”}.
Note: the keys are title-case and the values are uppercase.

Therefore the valid_state() function converts the passed values into title/upper case and checks it through the dictionary. This function acts as a boolean checker that sets the value of valid_state_check to True if the condition is satisfied.

Variables used:

state_name: state name/value to be checked through the function (String)
us_states: A dictionary imported from the state_city_database.py (Dictionary of Strings)
Input- state name/ value to be checked
Output- True/False (Boolean)

remove_null()

This function removes all the null and missing values from the data. Values like NA, null, and Nan are removed to avoid skewness in the results.

remove_spaces()

This function removes all leading and trailing spaces (if any) from the values. This ensures that the data is properly formatted and consistent, making it easier to work with and reducing the chances of errors in the code.

remove_outliers():

This function takes in the final scores/confidence values. 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 value 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.

Conditional Reasoning:

Based on the final values of the checkers: colname_check and valid_state_check, the get_sconfidence_score() assigns the following confidence scores:

Condition	Score	Reasoning
col_ckeck - True Valid_state_check - True	100	If the column name matches and the value is found in the state dictionary, it is certain that the value is a valid US State and is given a 100% score.
col_ckeck - True Valid_state_check - False	60	If the column name matches but the value is not found in the state dictionary, it could

		be possible that there is a spelling error/undetected missing value. The column name is given a higher priority hence the score is a 60%
col_ckeck - False Valid_state_check - True	40	If the column name does not match but the value is found in the state dictionary, the score is set to 40% for passing one condition.
col_ckeck - False Valid_state_check - False	0	If neither of the conditions is satisfied, it is certain that the value is not a valid US state and the score is set to 0%

Example:

Input:

col_name= “State”

col_values= [“california”, “Washington”, “TX”, None, “n/a”, “93403”, “FL”, “Belgium”]

After these values are passed through the plugin, these values will be set as scores

scores=[100.0, 100.0, 100.0, 60.0, 100.0, 60.0]

Note that the None and n/a values were removed by the remove_null() function

The outliers are further removed by the remove_outliers() function and the final score of 100.0 is returned after calculating the median of all the scores.

Implications for the Future:

In the future, we would like to handle international states. This means adding more databases of states all around the world or using an external API to check if a value is a valid State or not.