README - Data Analysis and Combination Probability Tool

Introduction

Welcome to the Association Pattern Discovery Program! This program is designed to analyze CSV files and uncover patterns of association among the data columns. This guide is tailored for users with basic technical skills and will walk you through every step needed to successfully run and navigate the program.

Overview

This Python program is designed to process CSV data, compute various probabilities, and export the results. It allows for analyzing combinations of data columns, calculating probabilities, entropies, chi-squared values, and more.

Requirements

Before starting, ensure you have the following:

- Python Installation: The program is written in Python. If you don't have `Python 3`, download and install it from python.org.
- Pandas Library: This Python library is essential for the program. Install it by running `pip install pandas` in your command line or terminal.
- CSV Data File: You'll need a `CSV` (Comma-Separated Values) file with the data you wish to analyze and it must be located where the `main.py` file is.
- Familiarity with Terminal or Command Line: Basic knowledge of how to navigate folders and run commands in a terminal or command line is required.

Installation Guide

Step 1: Downloading the Program

- Download main.py and this ReadMe file to a known directory on your computer.
- Ensure you also have a CSV file that you wish to analyze in the same directory.

Step 2: Handling the Zip File

- Download the provided .zip file.
- Extract the contents of the zip file to a directory of your choice. This directory should now contain main.py.

Running the Program

Option 1: Using an IDE

- Open your preferred IDE (e.g., VSCode, PyCharm).
- Open the directory where you downloaded main.py and the CSV file.
- Run main.py using the IDE's run functionality.

Option 2: Using Replit

- Go to Replit https://replit.com/@YousufKh/AssociationPatterDiscovery#main.py
- Click 'Run' to start the program.
- Navigate to '< >' under the program name to see all the other files including ReadMe file.
- Navigate to Exported to see the exported result.

Option 3: Using Terminal from the Unzipped Folder

- Navigate to the directory where you unzipped the .zip file.
- Open a terminal or command line window in this directory or navigate to this directory in terminal.
- Run the command `python main.py` to start the program.

Program Navigation and Usage

Initial Prompt

Upon starting the program, you will be asked if you have read the ReadMe file. The program will loop through this prompt until you respond with 'Yes'.

Loading the CSV File

Enter the path to the CSV file when prompted. Ensure you have the CSV file that you wish to analyze in the same directory of `main.py`. Enter the name exactly as it is as the Program is `case sensetive`.

Data Range and Support Threshold

Specify a data range (like 'A1:D10') for exactly which table you want to analyze in your CSV File. Ensure that the table starts with header name and included it. If not, the program will take the first row as header. Ensure all the other row only consists of Numbers / Integers. Program will crash otherwise.

Enter a support threshold value (e.g., 0.05). Choose your own threshold. Any invalid input will default it to 0.

Command List

table: Show the entire uploaded data table. ***header***: List all headers/column names. ***show***: Display probabilities for a specific column. ***perms***: Explore header combinations. ***viewdata***: View detailed data stored in the program. ***export***: Export combination probabilities to CSV files. ***quit***: Exit the program.

Features

- Load and process data from CSV files.
- Calculate column-wise probabilities.
- Generate and analyze combinations of data columns.
- Calculate joint probabilities, information measures, and chi-squared values.
- Export calculated data to CSV files.

Main Functions

CSVPreprocessor: Loads and processes CSV files. *ColumnProbability*: Calculates probabilities for each column in the data. *CombinationProbability*: Analyzes combinations of columns for joint probabilities and other statistical measures. *Export to CSV*: Exports the results of the calculations to a CSV file. *Main*: handles the input, commands etc.

Usage

Load Data: Start by loading your CSV file. You will be prompted to enter the file path and the desired data range (e.g., 'A1:D10'). ***Calculate Probabilities***: The program automatically calculates individual column probabilities and stores them. ***Explore Combinations***: Use the 'perms' command to explore header combinations. You can specify the order (e.g., '5' for up to order 5, '15' for only order 5). ***View Data***: The 'viewdata' command allows you to inspect the currently loaded DataFrame, probability results, header combinations, and combination probabilities. ***Export Data***: Use the 'export' command to export the combination probabilities to CSV files. You will be prompted to enter specific headers and the maximum order for export. Only rows with AssociationValidity = True will be exported. ***Additional Commands***: Use 'show', 'table', and 'header' commands to display specific column probabilities, the entire data table, and the list of headers, respectively. Check the Examples below to get a more profound understanding.

CSV File Processing

The program can process specific ranges within a CSV file. Specify the range in the format 'A1:D10' when prompted.

Probability and Statistical Calculations

Calculates column-wise probabilities. Analyzes combinations of columns for joint probabilities. Computes information measures, chi-squared values, and significance. Special calculation for combinations of order 2 and higher. *Only calculation upto Order of 6 is made because anything over is assumed to be irrelevent*

Examples

- Starting the porgram:

  ```
  Have you read the 'ReadMe' file yet?
  (Yes/No): Yes
  ```

- Load CSV: Make sure it the name matches exactly same as the CSV file you want to import. It is case sensetive. Here the provided CSV file is EncodedData.csv with data range A1:O155.

  ```
  Enter the path to your CSV file: EncodedData.csv
  Enter the desired data range (e.g., 'A1:D10'): A1:O155
  DataFrame loaded and processed successfully.
  Calculation in Progress ...
  ```

- Choose a threshold:

  ```
  Choose a proper support threshold. Such as 0.02 or 0.05 or 0.5 etc.
  Enter the support threshold (default 0): 0
  support threshold is set to: 0.0
  Calculating information measures and chi-squared values for combinations...
  ```

```
Calculation Complete.
Command: ('show', 'table', 'header', 'viewdata', 'perms',' export', 'quit')
Enter a command:
```

- table: prints the imported data table from the csv

```
Command: ('show', 'table', 'header', 'viewdata', 'perms',' export', 'quit')
Enter a command: table
0     Q28 Q29b S6 Q8 Q9 Q12 Q16 Q17 Q18 Q19 Q24 Q45 Q46 Q47 Q48
1       4    1  1  1  4   4   2   1   4   3   4  16   3   1   5
..    ...  ... .. .. ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..
154  2048    3  0  5  5   4   3   3   1   5   5  16   4   8   1
```

[154 rows x 15 columns]

```
- header:
```

Command: ('show', 'table', 'header', 'viewdata', 'perms',' export', 'quit') Enter a command: header ['Q28', 'Q29b', 'S6', 'Q8', 'Q9', 'Q12', 'Q16', 'Q17', 'Q18', 'Q19', 'Q24', 'Q45', 'Q46', 'Q47', 'Q48']

```
- Show:
```

Command: ('show', 'table', 'header', 'viewdata', 'perms',' export', 'quit') Enter a command: show Enter the column name to show probability: Q29b Q29b Count Probability 0 1 75 0.487013 1 2 42 0.272727 2 3 30 0.194805 3 4 7 0.045455

```
- perms:
prints the header combination of all the header names. Use 3 to get all the combination upto Order of 3 and 13 to get only the combinations of 3 headers.
```

Command: ('show', 'table', 'header', 'viewdata', 'perms',' export', 'quit') Enter a command: perms Enter the order for header combinations (e.g., '5' for up to order 5, '15' for only order 5): 12 Order 2 combinations: [('Q28', 'Q29b'), ... , ('Q47', 'Q48')]

```
- View Data: Enter a command: viewdata
This prints all the calculated data on the terminal.
```

Command: ('show', 'table', 'header', 'viewdata', 'perms',' export', 'quit') Enter a command: viewdata {'Q16': '3', 'Q18': '1', 'Q45': '1', 'Q46': '3', 'Q47': '4', 'Q48': '1', 'Count': 2, 'Probability': 0.012987012987012988, 'InformationMeasure': 6.840917988150577, 'ShannonEntropy': 5.7596499099457965, 'MaximumEntropy': 6.475733430966398, 'ChiSquaredBy2N': 0.7314600792340988, 'Significance': 61.792566405354115, 'MI - Chi_sqr_2N': -54.951648417203536}, ... {'Q19': '5', 'Q24': '5', 'Q45': '8', 'Q46': '3', 'Q47': '0', 'Q48': '5', 'Count': 1, 'Probability': 0.006493506493506494, 'InformationMeasure': 11.66690044567618, 'ShannonEntropy': 5.7596499099457965, 'MaximumEntropy': 6.475733430966398, 'ChiSquaredBy2N': 10.550404243108684, 'Significance': 808.1378816919503, 'MI - Chi_sqr_2N': -796.470981246274}]

```
- Export Data: where threshold of 0 was selected.
```

Enter a command: export Enter headers to export (separated by commas): Q28,Q29b,S6,Q8,Q9,Q12,Q16,Q17,Q18,Q19,Q24,Q45,Q46,Q47,Q48 Enter the maximum order to export (2-6): 6 Exported data to F:~\AssociationPatterDiscovery\Exported\Q28_Q29b_S6_Q8_Q9_Q12_Q16_Q17_Q18_Q19_Q24_Q45_Q46_Q47_Q48_uptoOrder6.csv Command: ('show', 'table', 'header', 'viewdata', 'perms',' export', 'quit')

```

### Understanding the Output CSV File

- Combination: Represents a unique combination of column headers from the original dataset. It shows which columns are being analyzed together for patterns of association.
- Q29b, ... , Q48, etc: These are specific column headers from your original CSV file. Each represents a distinct attribute or variable in your dataset.
- Count: The frequency count of each unique combination of values found in the columns specified in the "Combination" field.
- Probability: Joint Probability of that unique combination. The likelihood of each unique combination occurring within the dataset. It is calculated as the count of the combination divided by the total number of observations.
- ShannonEntropy: A measure of the uncertainty or randomness in the probability distribution of the combinations. Higher entropy indicates greater randomness. This can be derived as `E'`
```

- MaximumEntropy: The maximum possible entropy for the given combination, indicating the highest possible level of disorder or randomness in the distribution. This can be derived as `E^`
- InformationMeasure: This quantifies the amount of information obtained from observing the particular combination. It is calculated using the joint probability and individual probabilities of the elements in the combination. This can be derived as `MI`
- ChiSquaredBy2N: A statistical measure calculated for each combination, representing the chi-squared value divided by twice the number of observations (2N). It's used to test the independence of variables in the combination. This can be derived as `(λ)^2 /2N`
- Significance: A value indicating the significance of the observed combination. It considers the probability, entropy measures, and chi-squared values. For combinations of order 2, it is equal to ChiSquaredBy2N. This can be dervied as `(λ)^2 /2N` for order of 2 and `(1/jntPr){(λ)^2 /2N}^{(E'/E^)^(Order/2)}` for combination that is more than order of 2.
- MI - Chi_sqr_2N: This represents the difference between the 'InformationMeasure' and 'Significance' for each combination. It gives a sense of how the information measure compares to the significance value.

Notes

Ensure the CSV file path and data range are correct. The export functionality allows for selecting specific headers and the maximum order of combinations to be included in the export. Ensure your CSV file is correctly formatted and accessible at the provided path. The tool handles a range of data analysis tasks, but it's primarily focused on association patterns and their significance.