Seenopsis

Meytal Avgil Tsadok

Link to Git: https://github.com/meytala/seenopsis

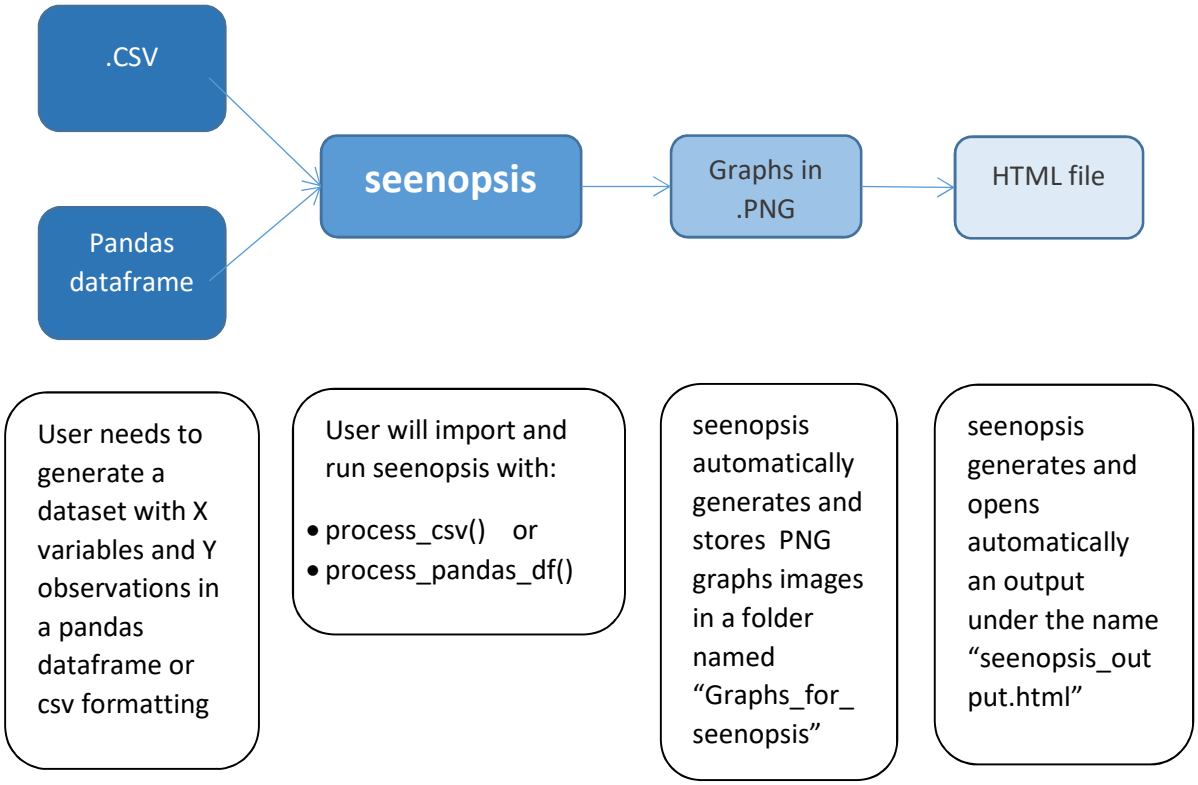
INTRODUCTION

seenopsis is a tool aiming to aid first exploration and visualization of available variables in a giving dataset. seenopsis centralizes the main important features of the different variables in a structured visualized approach.

TERMINOLOGY

- **Dataset** - a collection of data, set in a single table, where every column of the table represents a particular variable, and each row corresponds to a given observation.
- **Variable** - a symbolic name associated with a value and whose associated value may be changed.
- **Value** – a property assigned to a variable.

ARCHITECTURE



DATASET STRUCTURE

To use seenopsis, structure your dataset with the different variables as columns and observations as rows.
The following are required:

1. Each variable in the dataset should be placed in its own column
2. Each observation should be placed in its own row
3. Each value should be placed in its own cell
4. The first row should contain the name of the variables
5. Your dataset should not have a prefix/title within the dataset

USE CASE

seenopsis is intendent to be used by anyone who wants to have a first exploration of dataset's variables. In version 1.0.1, seenopsis users will choose one of the following options, based on the formatting of the dataset:

seenopsis.process_pandas_df() - for datasets that are in a *pandas* data structures (python):
After importing seenopsis, simply run this command, passing the name of the dataset.
After executing the *seenopsis.process_pandas_df()* command , a new html tab with the dataset's seenopsis will be opened in your default internet browser.

seenopsis.process_csv()- for datasets that are saved as a csv file**:**
If your dataset is not in a pandas dataframe (for example you are using R or SQL environments), simply convert it to a csv file and use seenopsis.process_csv(). seenopsis version 1.0.1 can read csv files that were encoded using utf-8, ANSI, ISO-8859-1 and ISO-8859-8.
After importing seenopsis and executing the seenopsis.process_csv() command, a new dialog window will be open and the user will have to point the path for the dataset saved as csv. Once the user choose the file and click open, a new tab with the dataset's seenopsis will be opened in the default internet browser.

REQUIREMENTS AND DEPENDENCIES:

In order to execute seenopsis the following libraries are needed:

- pandas
- numpy
- matplotlib.pyplot
- webbrowser
- tkinter.filedialog (askopenfilename)
- os
- sys

Additionally, you should have an internet browser installed on your computer (for example chrome or explorer). seenopsis will be better presented in chrome.

ADDITIONAL INFORMATION

While running seenopsis, a new folder named “Graphs_for_seenopsis”, will appear in the working directory. This folder is essential for seenopsis table output.

seenopsis OUTPUT

The seenopsis output is an html file containing a table, added to the working directory (as “seenopsis_output.html”).
The html table displayed automatically at the end of the processing.

seenopsis OUTPUT LAYOUT

In the seenopsis header you will file information on your dataset structured in the following format:
“The file you are investing has XX variables for YYY observations.
This is the seenopsis of your file: “

Follow the header, you will see the seenopsis information structured in a rolling table. The table contains 6 columns:

- **Variable Name:** the name of the variable explored in the dataset
- **Type:** the type of the variable explored
  Potential types available:
  - Single Value – one unique value, not including null
  - Binary Variable (text/date based) – two distinct values of a string or a date, not including null
  - Binary Variable (integer/float based) - two distinct values, of an integer/float values (i.e two distinct numbers), not including null
  - Categorical Variable (text/date based) - between 3 to 10 unique text/date values (not including null)
  - Categorical Variable (integer/float based) - between 3 to 10 unique integer/float values (not including null)
  - Continuous variable (int/float) – integer/float values with more than 10 unique values
  - Text/Date variable – a text/date with more than 10 unique values or other object types that are not integer/float
- Graphic Representation: varies based on the type of the variable
  - Single Value – horizontal bar chart
  - Binary Variable (text/date based) – horizontal bar chart

  - Binary Variable (integer /float based) - horizontal bar chart
  - Categorical Variable (text/date based) - horizontal bar chart
  - Categorical Variable (integer /float based) - horizontal bar chart
  - Continuous variable (integer /float based) – histogram
  - Text/Date variable – horizontal bar chart, only top 10 are presented.
- Basic Statistic: based on type of variable
  - Single Value – no statistics
  - Binary Variable (text/date based) / (integer/float based) – name and percentage of each value count
  - Categorical Variable (text/date based) / (integer/float based) - number of unique values
  - Continuous variable (integer /float based – minimum value (min), maximum value (max), mean ± SD, median (IQR (25%, 75%))
- Missing: number of missing values and percentage. If 0, indicates “No missing values”.
- Outliers: only in continuous variables. Presents the number of outliers, based on extremities in a distance of 3 IQRs from the median. If the IQR is equal to zero, outliers will not be analyzed.

EXAMPLES:

For a dataset formed in a pandas dataframe:

```
import seenopsis
seenopsis.process_pandas_df(name_of_dataset)
```

For a dataset formed as a csv:

```
import seenopsis
seenopsis.process_csv()
```