**seenopsis**

Link to Git: XXXXXX

**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**

.CSV

Pandas dataframe

**seenopsis**

Html file

Graphs in .PNG

User need to generate a dataset with X variables and Y observations in a pandas dataframe or csv formatting

User need to provide the name/location of the dataset

Generate and store automatically in a file named “Graphs\_for\_seenopsis”

Generate and open automatically under the name “output\_seenopsis.html”

**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)
* seenopsis.process\_csv() - for datasets that are saved as a csv file

**seenopsis.exe/IMPORTING SEENOPSIS**

**seenopsis.process\_pandas\_df()**

After importing seenopsis, simply run this command, passing the name of the dataset.

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():**

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, 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 in your browser 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

Additionally, you should have an internet browser installed on your computer (for example chrome or explorer). seenopsis is better present 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 “output\_seenopsis.html”).

The html table displayed automatically at the end of the processing.

In the seenopsis header you will file information on your dataset (number of observations and variables), followed by a 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 based) - two distinct values, of an integer 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 based) - between 3 to 10 unique integer values (not including null)
* Continuous variable (int64) – integer values with more than 10 unique values
* Continuous variable (float64) – 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 int64 or float64
* Graphic Representation: varies based on the type of the variable
* Single Value – horizontal bar chart
* Binary Variable (text/ate based) – horizontal bar chart
* Binary Variable (integer based) - horizontal bar chart
* Categorical Variable (text/date based) - horizontal bar chart
* Categorical Variable (integer based) - horizontal bar chart
* Continuous variable (int64) – histogram
* Continuous variable (float64) – 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 based) – percentage of each value count
  + Categorical Variable (text/date based) / (integer based) - number of unique values
  + Continuous variable (int64) / (float) – minimum value (min), maximum value (max), mean ± SD, median (25% q, 75%q)
* 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.

**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()