**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 - xxxxxxx

Dataframe - xxxxxxx

Variables - xxxxxxx

Values – xxxxxxx

IQR – Intra quartiles range

SD – standard deviation

**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 dataset’s formatting:

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

use process\_csv() – for datasets as a csv file

use process\_pandas\_df() – for datasets as pandas dataframe

**seenopsis.process\_csv():**

After importing seenopsis and choosing this option, a new dialog window will open and the user will choose the path for the csv dataset.

Once the user choose the file and click open, a new tab with the dataset’s seenopsis will be opened in an internet browser.

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

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

Executing the command , a new tab with the dataset’s seenopsis will be opened in an internet browser.

**Requirements and Dependencies:**

In order to execute seenopsis the following libraries are needed:

pandas

numpy

matplotlib.pyplot

webbrowser  
tkinter.filedialog   
os

Additionally, for displaying a proper html table, there is a need to copy the file “bootstrap.min.css” to the seenopsis working directory. The file can be obtained from [here](https://github.com/meytala/Final_project/blob/master/bootstrap.min.css).

**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.

The prefix of the table contain information about the dataset, including number of records and variables.

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\date 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”.
* Outliers: only in continuous variables. Presents the number of outliers, based on extremities in a distance of 1.5 IQR from the median.