Summarizer

Background In data analysis we are operating with large dataframes, which can contain different types of data, such as binary, numeric, datetime and classification variables.
Manual analysis of such dataframes is not efficient, and it could be very helpful to have some statistical overview of the dataframe.

Task description

The program should take pandas dataframe as input, iterate through each of the columns in the dataframe and based on column datatype, create summary statistics for each, and print them out in a table.
Calculated summary could include following items:

- column type
- min, max
- mean, median, mode
- percent of zero rows
- variance and standard deviation
- interquartile range and coefficient of variation
- number of distinct values

Feel free to add any other statistical measures if you find them useful.

Input

Pandas dataframe and options to customize output, for example:

- output_type - markdown, html or xlsx
- out filename
- other options if needed

Use IRIS dataset (https://archive.ics.uci.edu/ml/datasets/iris) as an example input.

Output

A markdown, html or xlsx report with summary statistics

What needs to be done

- Create python project
- Use OOP approach
- Test the solution
- Add README file
- Add requirements.txt file
- Do not ‘overengineer’ your solution Technical requirements
- Python 3.9 or higher
- Pandas

You’re allowed to use Open Source libraries and frameworks of your choice. We must be able to run your program simply by using Python and pip. We won't be able to evaluate your solution if it depends on proprietary libraries we don't have licences for.

Time

8-16 hour