report

December 1, 2022

0.1 Statistical Data Analysis II - Project I

0.1.1 1. Exploration

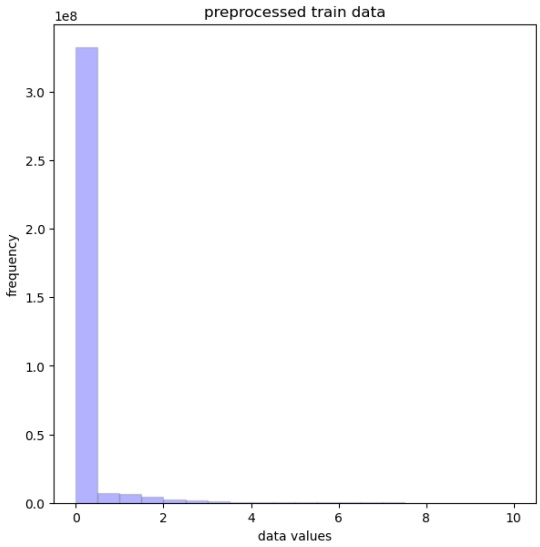
I used the read_h5ad() function from the scanpy package to load the datasets. The number of observations and variables in the loaded datasets is summarized in the following table:

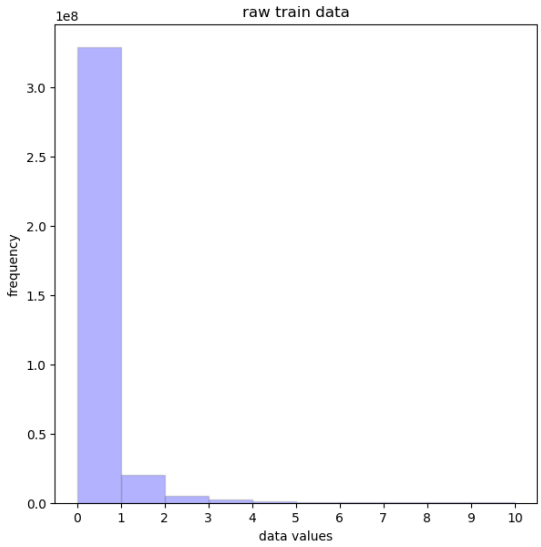
	Number of observations	Number of variables
train dataset	72208	5000
test dataset	18052	5000

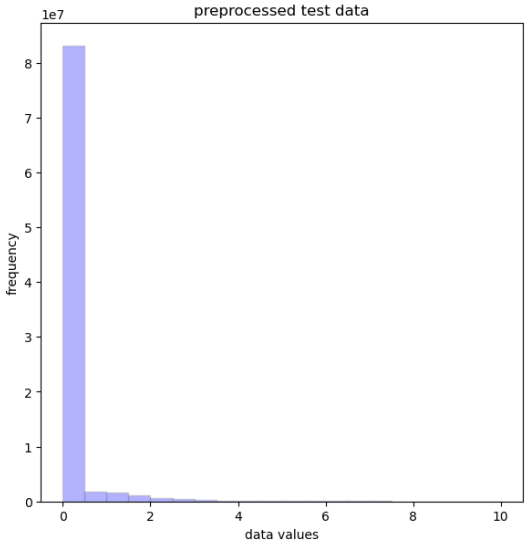
In the table below, I have presented the statistics for the loaded datasets.

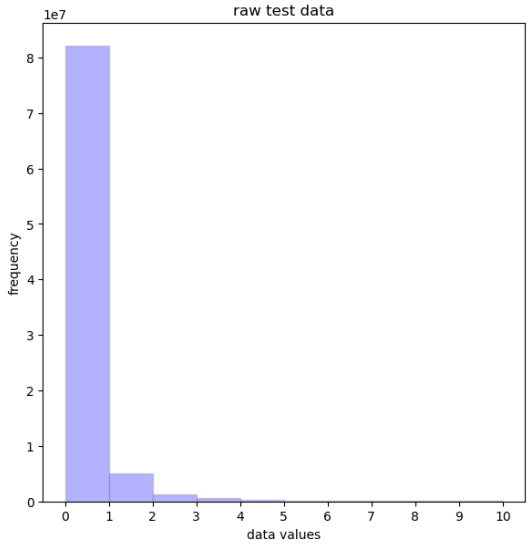
	min	max	mean	median
preprocessed train data	0.00	21,078,940.00	3.43	0.00
raw train data	0.00	35,451.00	0.44	0.00
preprocessed test data	0.00	21,078,940.00	3.66	0.00
raw test data	0.00	35,451.00	0.45	0.00

Below I have presented the histograms for the loaded datasets. Given the very large difference between the maximum and minimum values in the preprocessed and raw data, and the low values of the mean and median, I decided to limit the x-axis on the histograms to increase their readability.


preprocessed train data
1e8
frequency
data values


raw train data
1e8
frequency
data values


preprocessed test data
1e7
frequency
data values


raw test data
1e7
frequency
data values

[ ]:

[ ]:
```
i=5
```