Nile University

Machine Learning

Assignment 1 (Text Data analysis)

Eman Moustafa Ismail

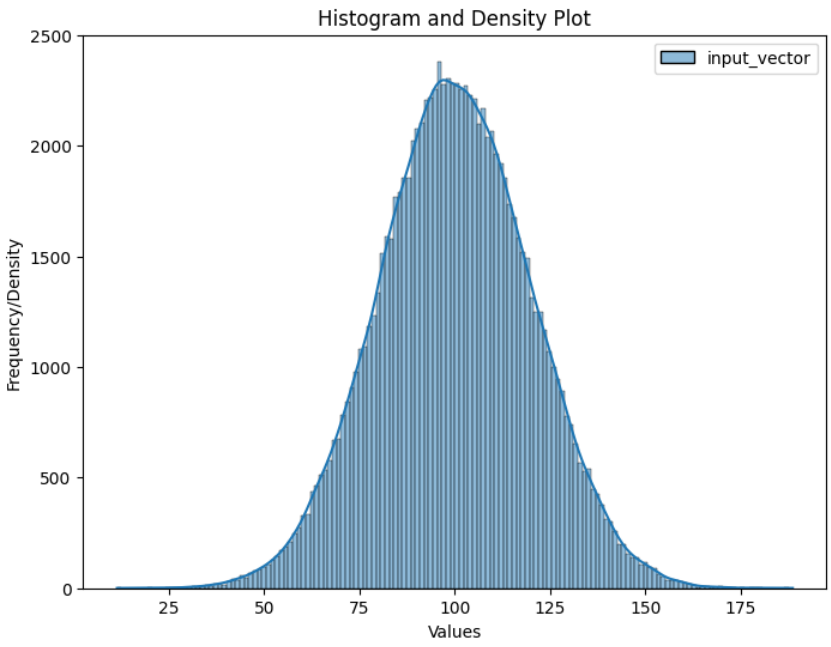
28th November, 2023

Data Statistical Measures Analysis:

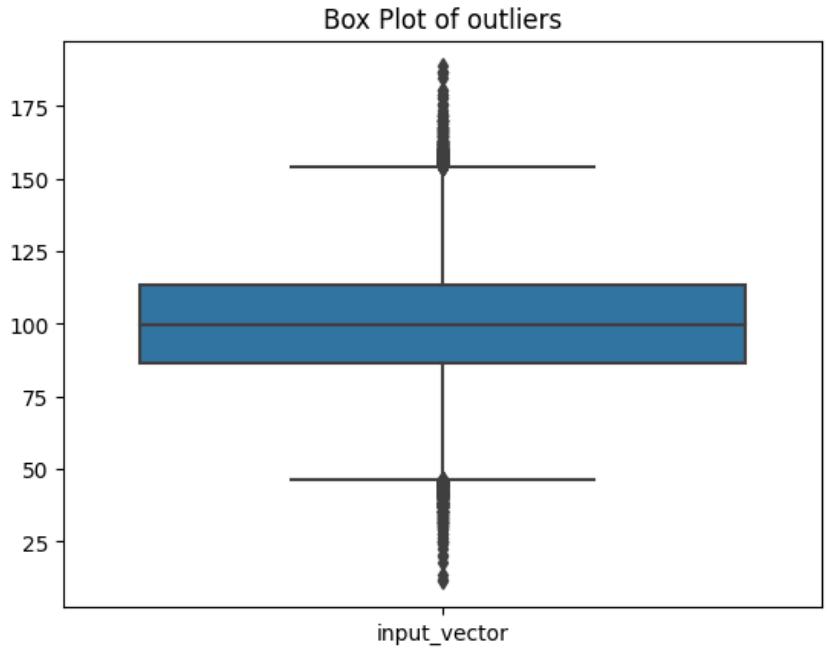
- The mean = 100, whereas the Median = 99.8 which indicates that the data has a symmetrical distribution and the data is fairly distributed on both sides of the mean.
- Given the minimum, Maximum and Standard deviation, I can claim that the change rate in the data is not bad. However, determining if the Variance is large or not is very subjective and depends hardly on the issue that needs to be tackled, other data inputs(if exist). Thus analyzing the variance should be always done in context.

Skewness and kurtosis Analysis:

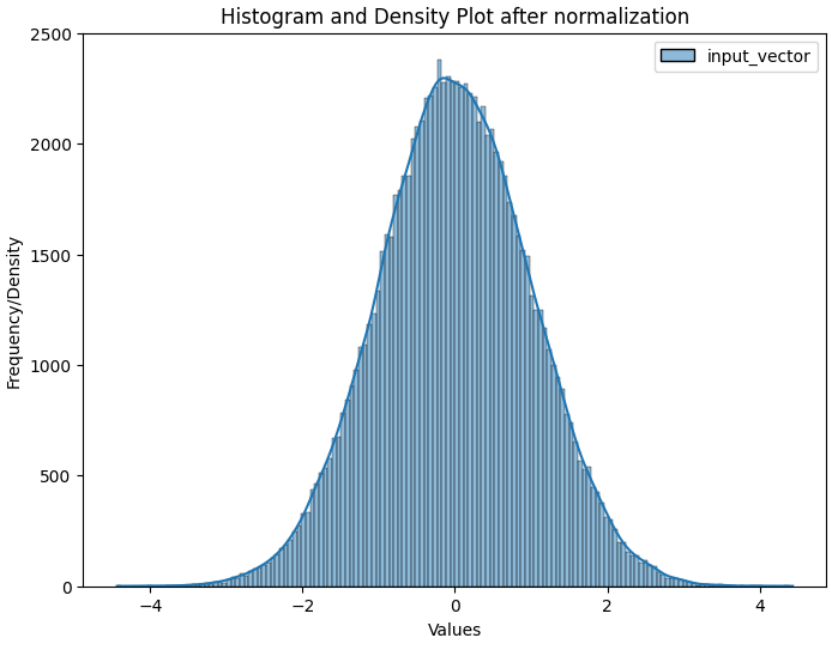
- The Skewness and Kurtosis is almost zero. This indicates that the data is normally distributed and generally symmetric. This proposal could be backed up by the fact mentioned in the previous analysis that the mean is almost equal to the median.
- Visualizing the data is also showing the normality in the distribution



Outliers and Data Transformations:



- Out of the 100k points there are ~700 outliers. However, due to having a skewness almost equal to zero, and fairly distributed data outliers are not considered as red flags.
- Significant transformation is not needed in our case. However we can simply apply a standardization function to get a data in the standard format.
- Applying data transformation should also take into consideration the problem context, thus in our case my judgment could be inaccurate.


Histogram and Density Plot after normalization
input_vector
Frequency/Density
Values
2500
2000
1500
1000
500
0
−4
−2
0
2
4