PCA for dimension reduction

Background

A complex modern semi-conductor manufacturing process is normally under consistent surveillance via the monitoring of signals/variables collected from sensors and or process measurement points. However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information, irrelevant information as well as noise. It is often the case that useful information is buried in the latter two. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then feature selection may be applied to identify the most relevant signals. The Process Engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learning and reduce the per unit production costs. It's a larger dataset than we've used so far, and it has a lot of features—590. Let's see if we can reduce that. You can find the dataset and more information at http://archive.ics.uci.edu/ml/machine-learning-databases/secom/.

Task : Implement PCA for dimension reduction. If we want to capture 99% of variance, how many principle components we need? Complete the following table.

Principle component number	% variance	% cumulative variance
1		
2		
3		
…		
k		

Hint on data preprocessing: there are lots of NAN in the dataset. For each of them, replace it using its corresponding feature's mean value (mean calculated using non-NAN data).

1. **DO NOT SUBMIT DATSET!**
2. **We run plagiarism check for submitted code. Please don't look for solutions online.**
3. You will **not receive any credit if you directly use off-the-shelf machine learning tools**. Mathematic computation tools such as Numpy and other basic tools.
4. Submit your code (**.py file**) and your report (**no more than 1 page in .pdf**). Not following the submission requirement, e.g., **code in txt/pdf file, exceeding report page limit, etc. will receive a credit penalty.**