Assignment 5: PCA implementation
Ishmael Contreras
EEL 4930: Introduction to Machine Learning

Principal Component Analysis:
Is a data analysis technique that attempts to reduce the dimesionality of the dataset to make it easier to process and handle in large computations. PCA naturally is set to minimize as much of information loss as possible. The reason we use PCA is to apply mathematics and statistics that are too difficult to make sense of in large dimensions. PCA is very applicable when the data obtained has enough correlation between the variables, otherwise the eigenvectors will be too dissimilar and not form a distinctive vector space.

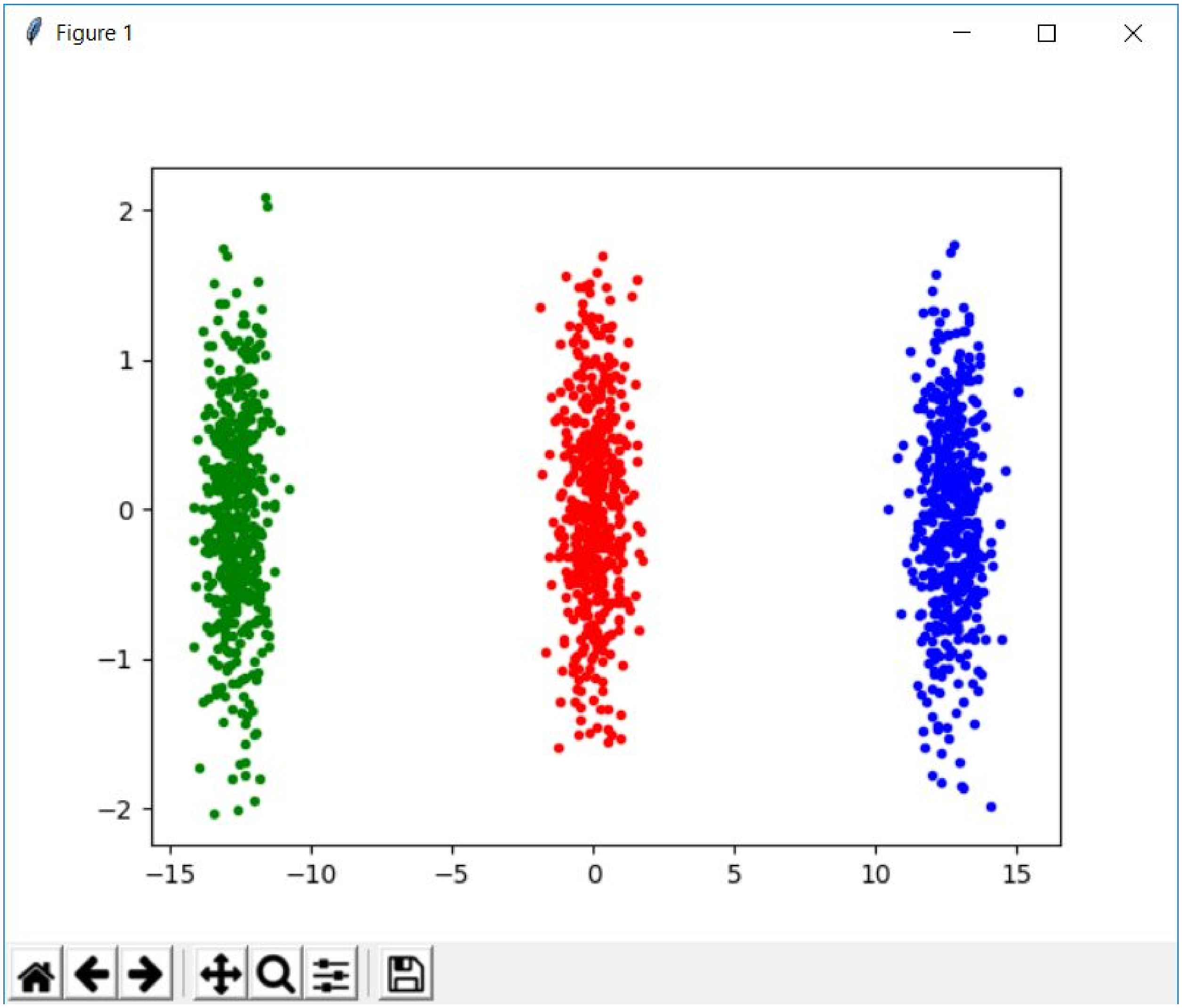

Figure 1-1: Dimension reduction of G1 data set from 10 dimensions to 2, the data is spread out and discernable so that it can be used without overlapping.

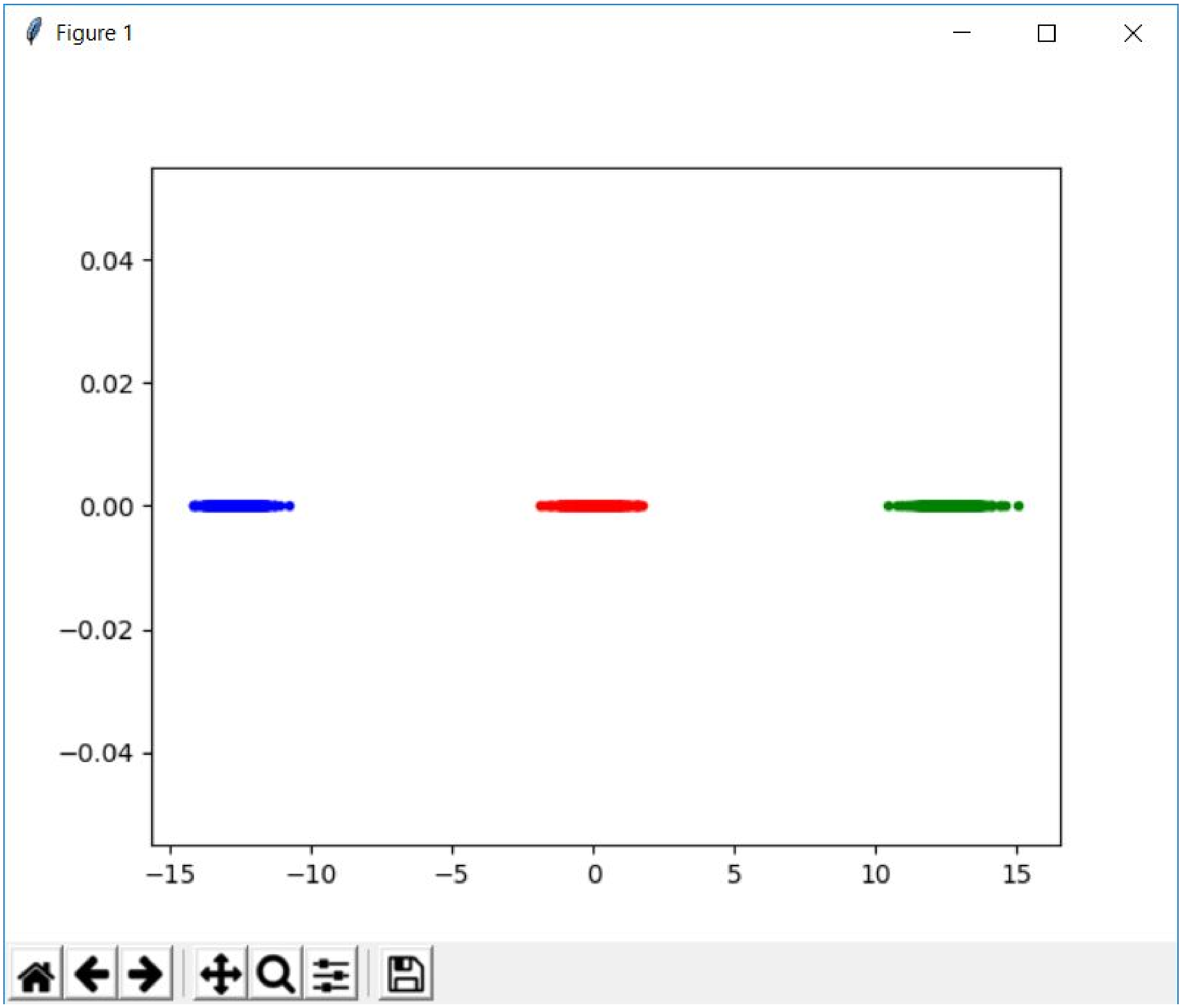

Figure 1-2: G1 in dimension one also is partitioned in a way that the data does not overlap, so even in 1D the largest eigenvalue is strongly representative of the variance in the data.

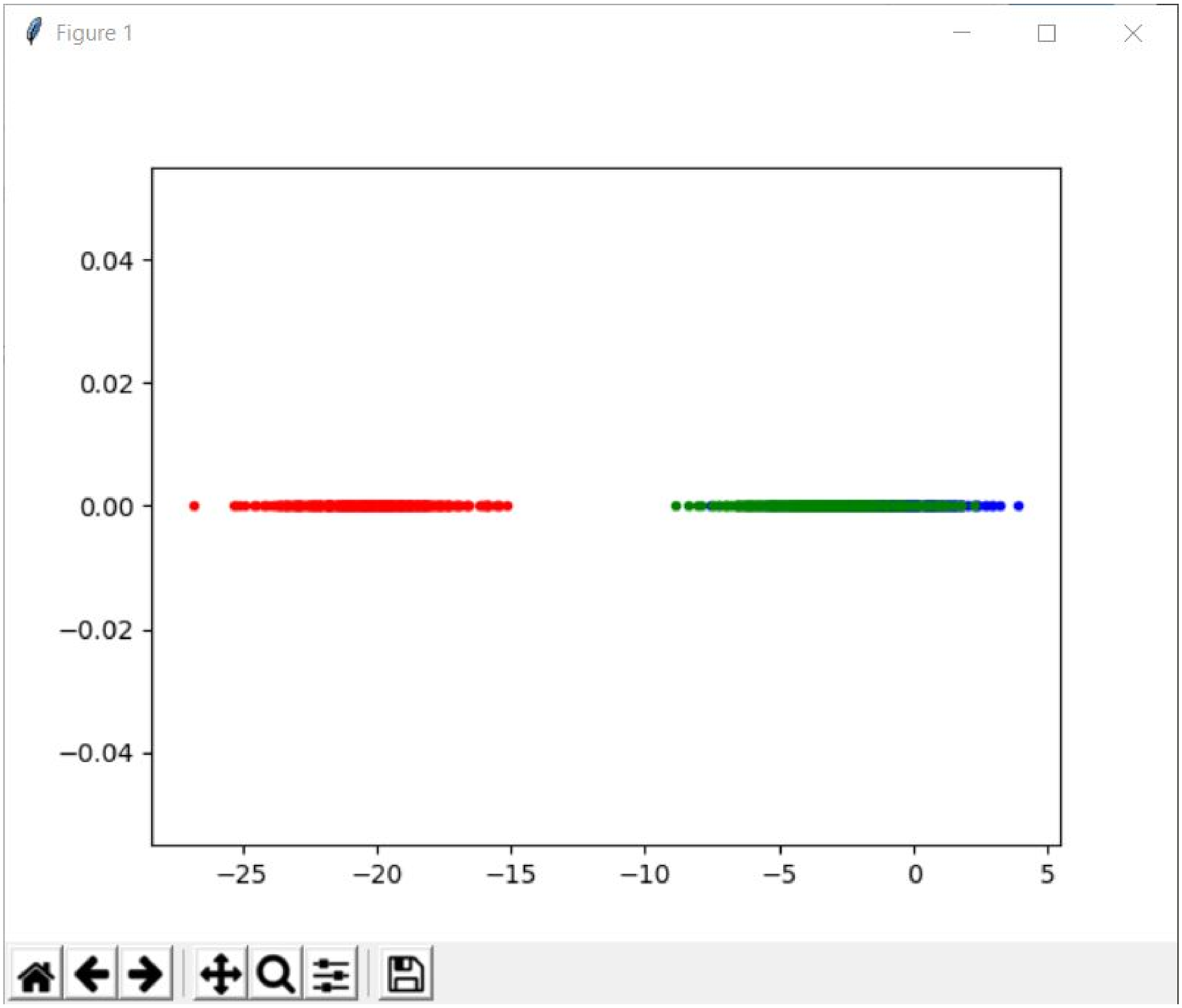

Figure 2-1: The PCA does not do a good job in differentiating the blue and green gaussian clouds. I would guess that the correlation between them is not strong enough for PCA do separate them in smaller dimensions.

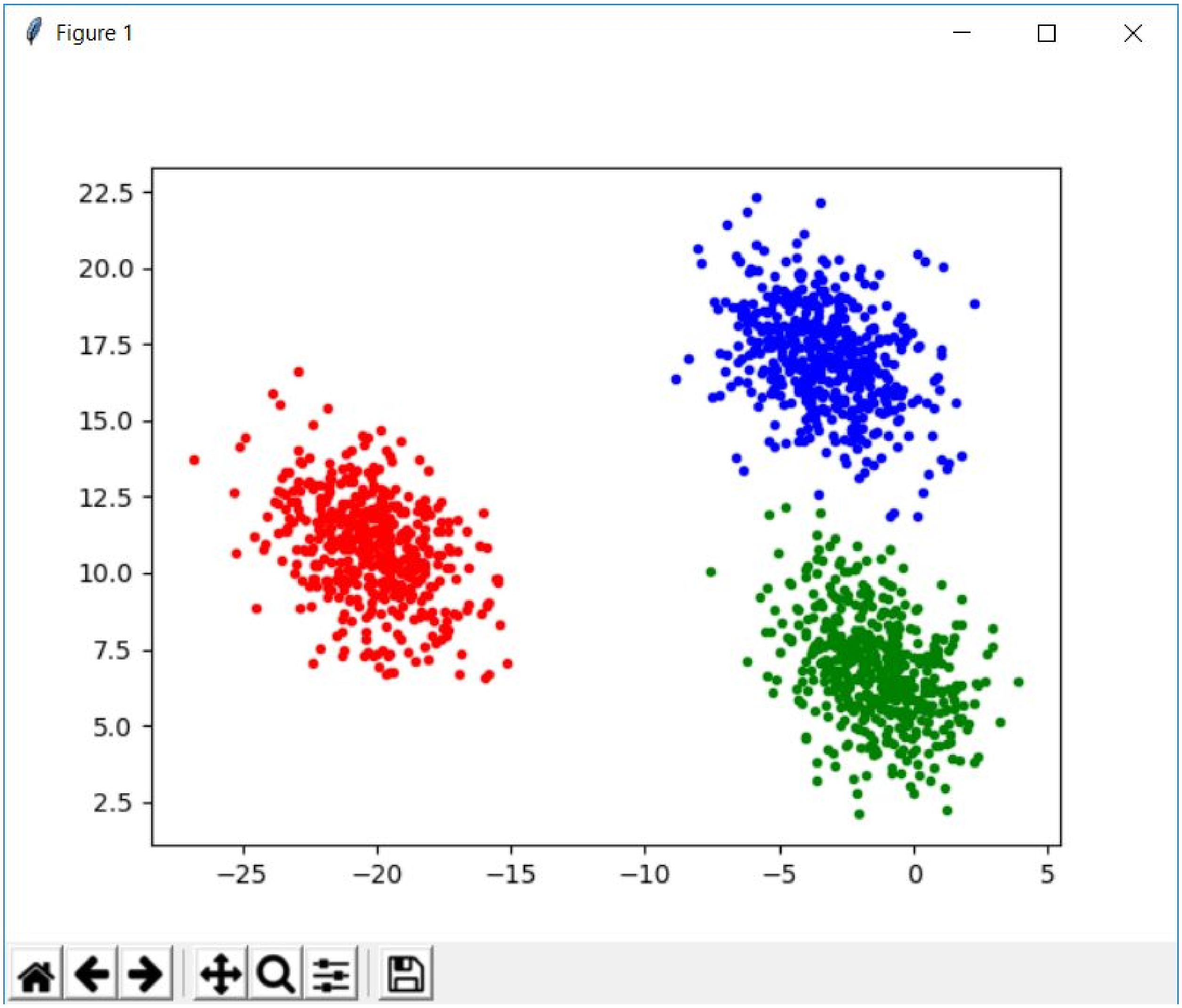

Figure 2-2: As i mentioned in the previous figure caption, The blue and green seem to be closely uncorrelated here in 2D because the scatter plot reveals some crossing of the data points around y ~11.

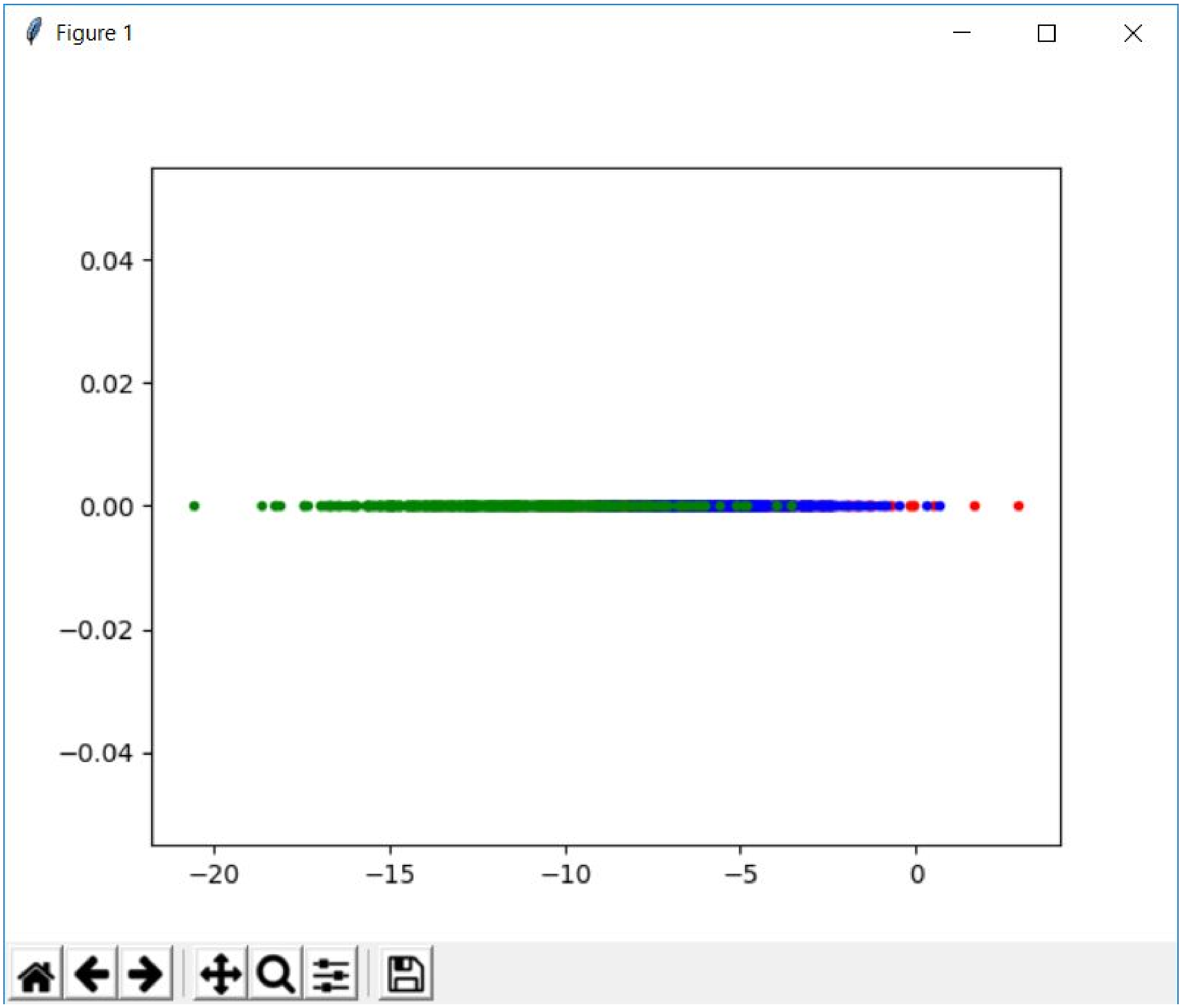

Figure 3-1: The green gaussian cloud seems to consume the the rest of the data in 1D. I can conlcude that the first eigenvector does not contain a large percentage of the entire data variance.

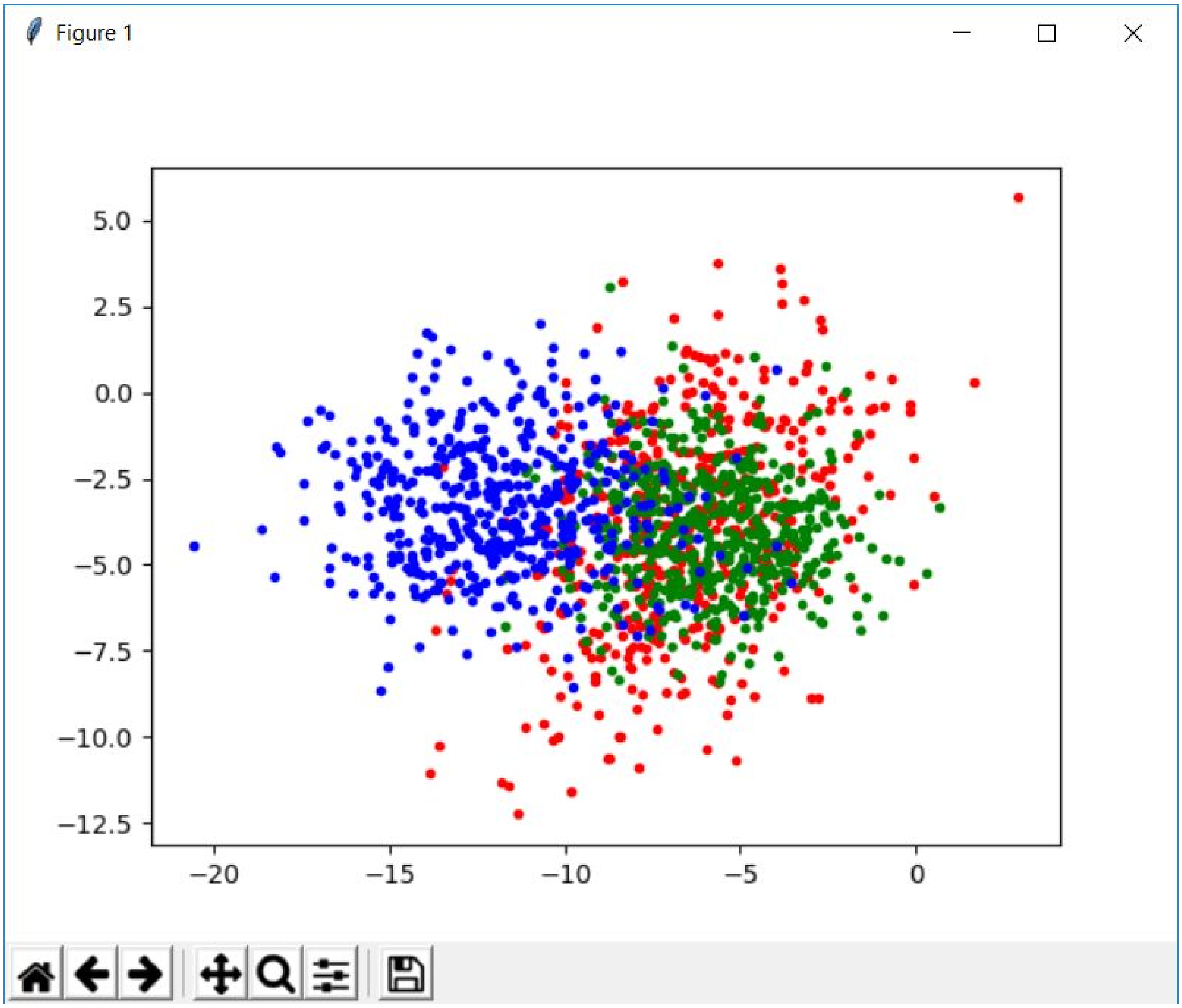

Figure 3-1: The same can be observed in the 2 dimensional space: The green data seems to share many regions with the blue and and red data.

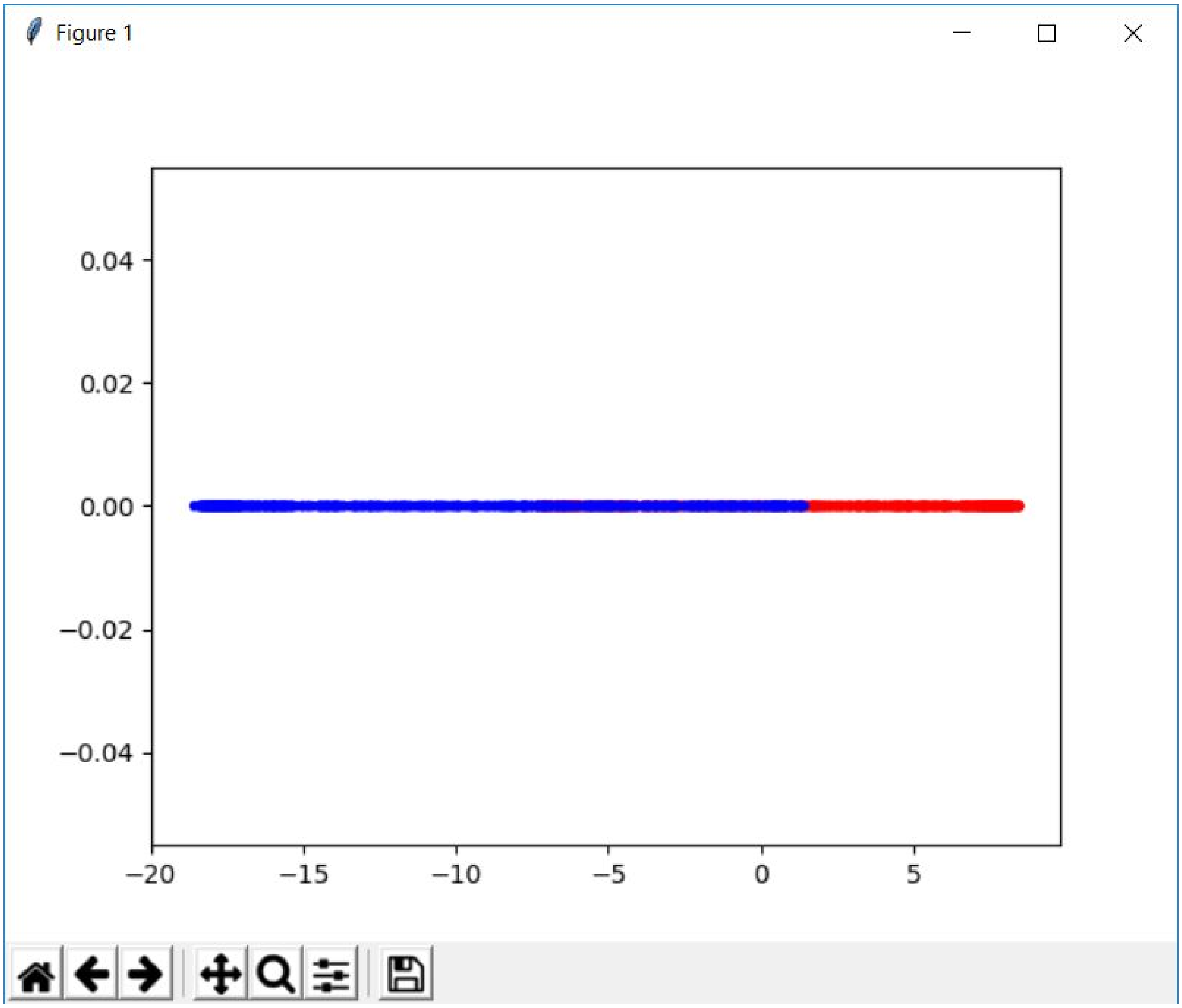

Figure 4-1: The “blue moon” coincides with most of the data of the “red moon”. This can also be seen in the original data plot of the “moons” that in an arbitrarily drawn line through the data that contains the percentage of overlap, describes the scalar value shown in the 1D projection.

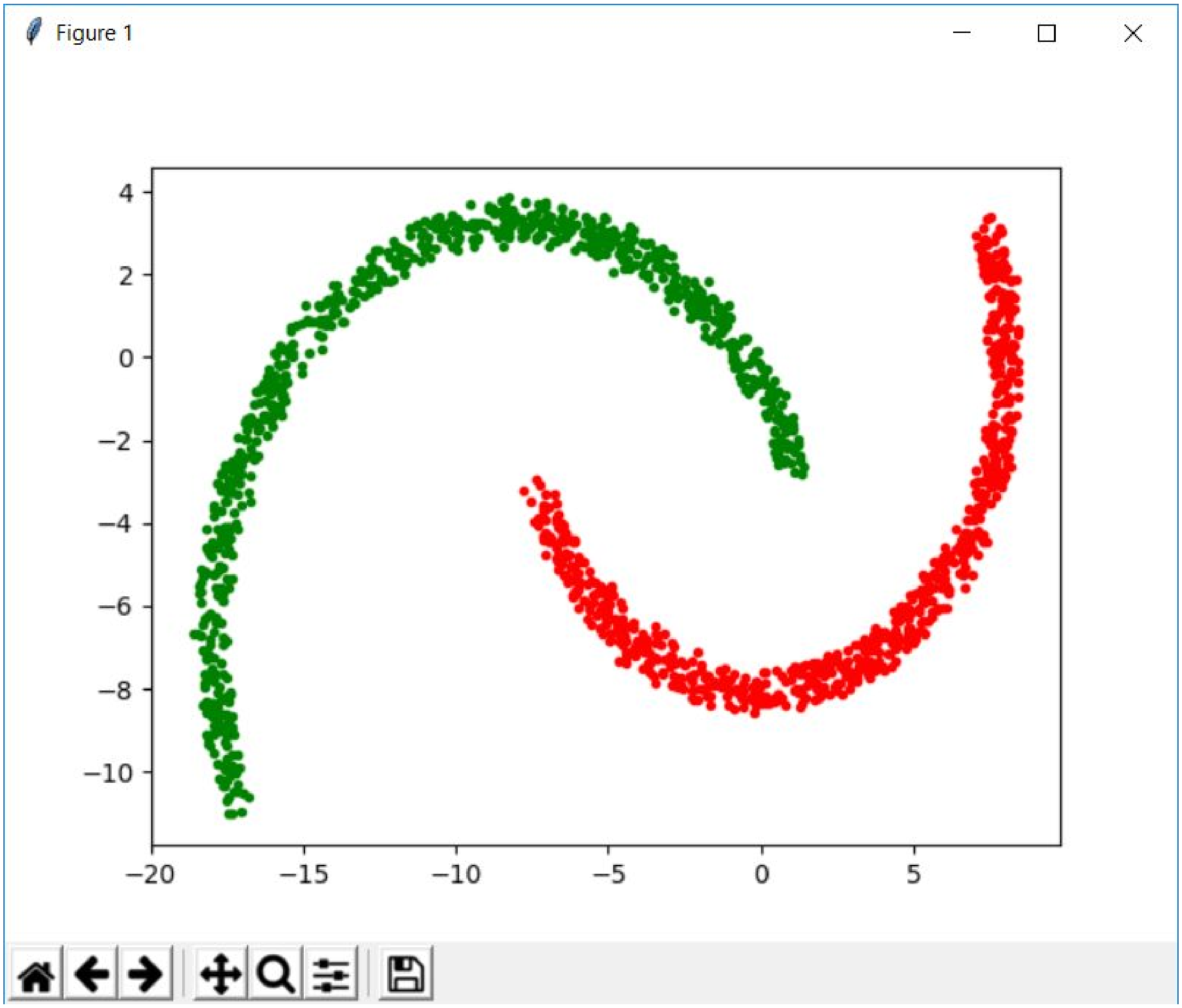

Figure 4-2: This 2D basis gives the data the contours represented in the original data which means that PCA has done a good job in preserving some of the original variance.

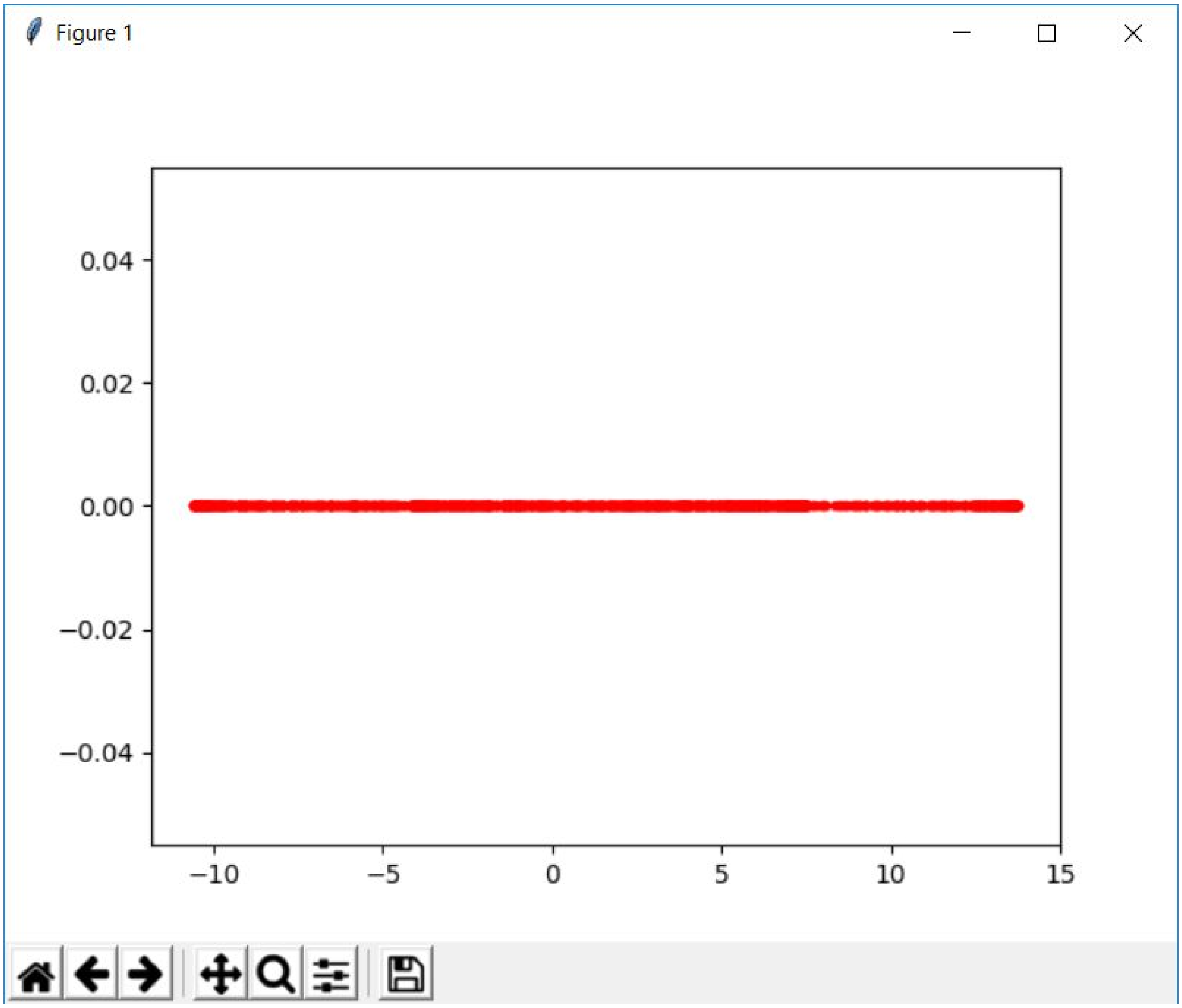

Figure 5-1: The swiss roll plot for 1D is not descriptive in telling us anything of the original data. Therefore, it is not advisable to project to only 1D.

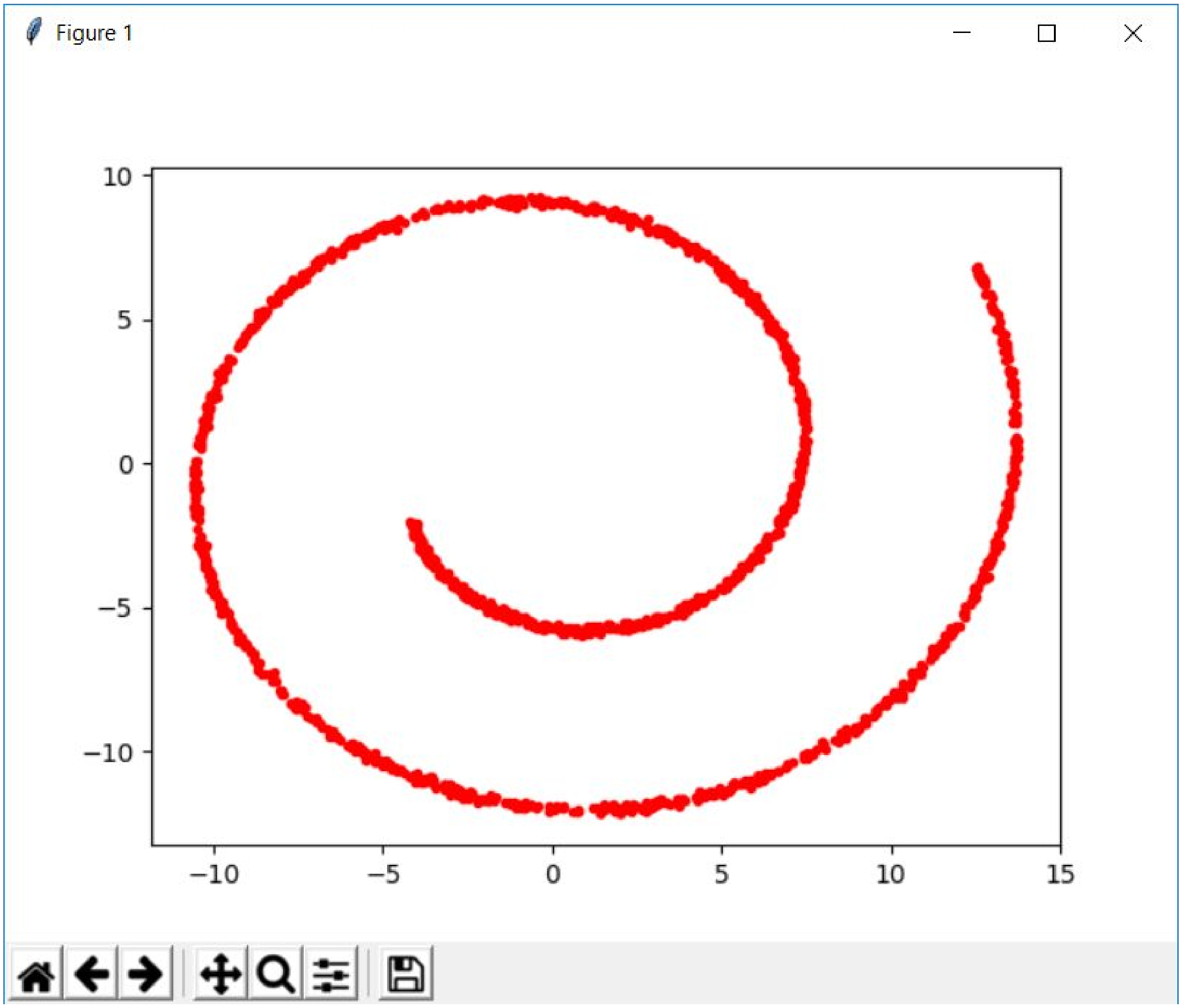

Figure 5-2: The shape of the original data is represented as a “cross-section” of the swissroll data set so we can see that PCA preserves some features of the original set.

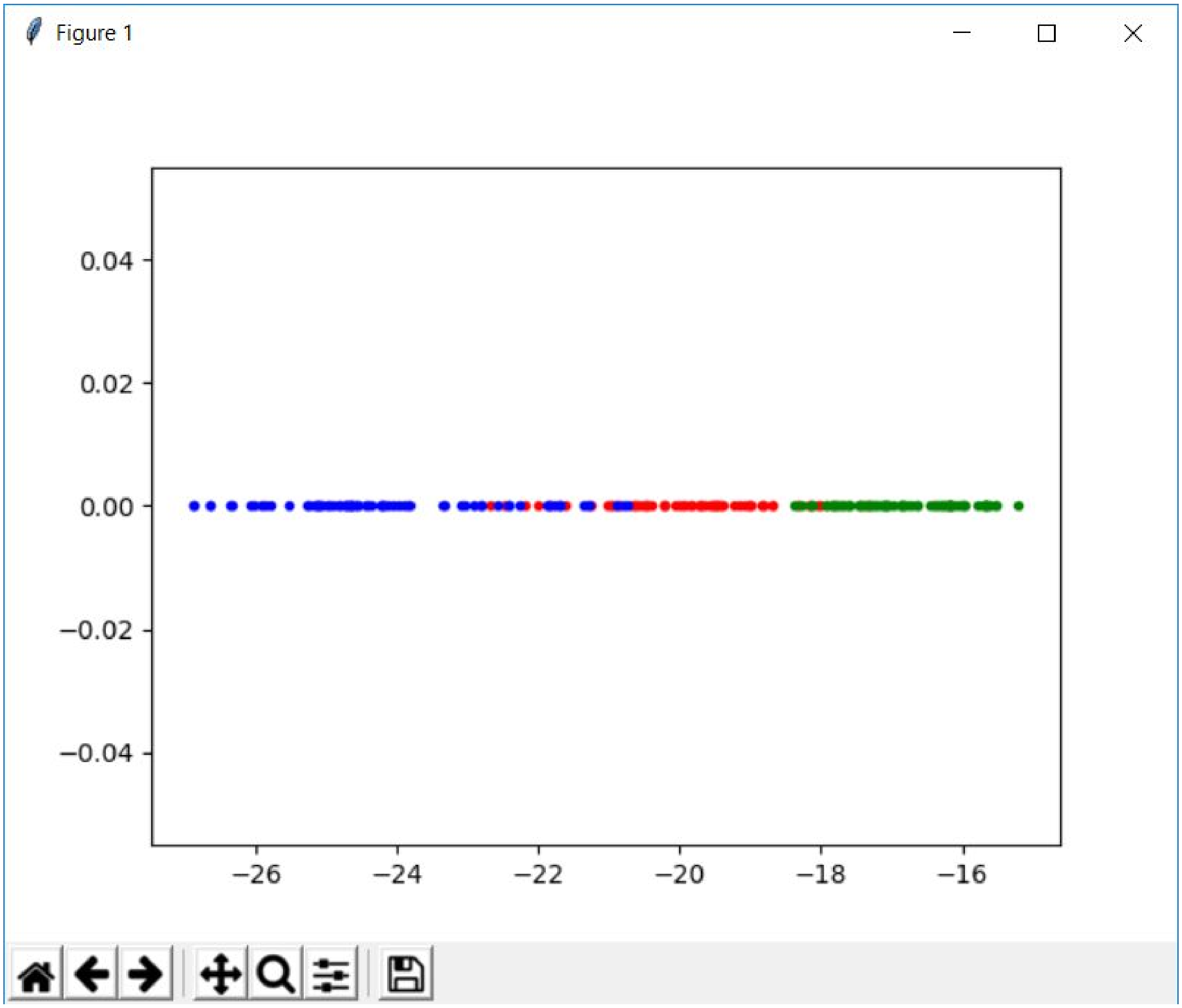

Figure 6-2: Each color is visibly represented, but some colors are overlapped by other colors, which tells us that the correlation matrix is correlated but not enough to form discrete lines.

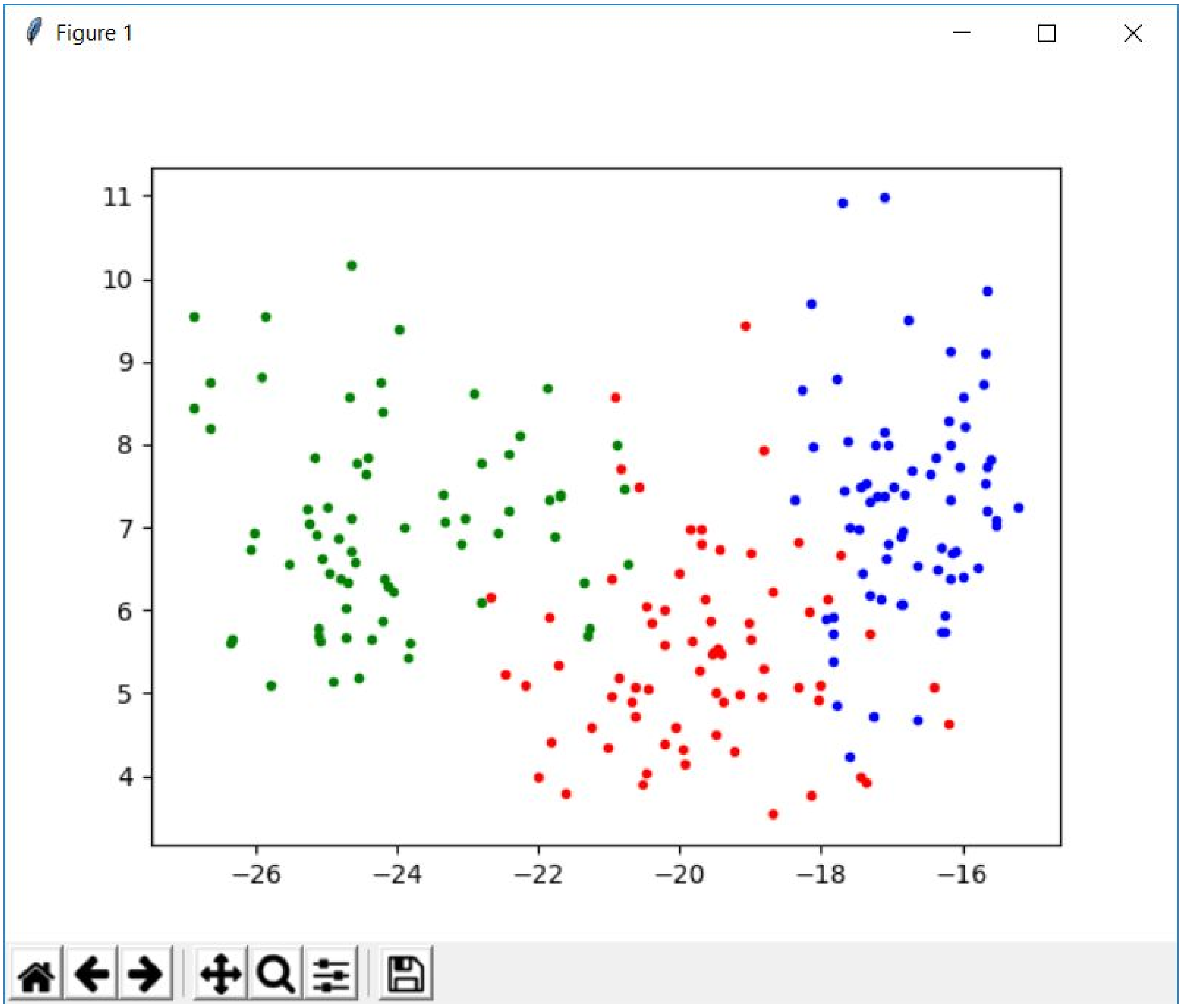

Figure 6-2: The seeds data in 2D is more spread out and is almost clustered enough to perfectly differentiate. If PCA were to project to the 3rd dimension, the data would look better. However, The 2D reduction is able to tell us enough about the data spread that we could use it.