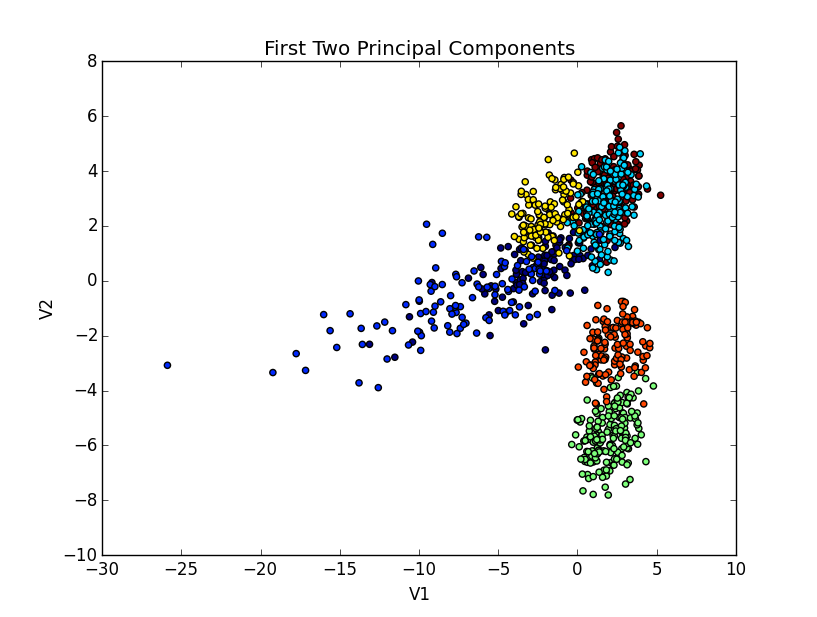
**Project 4**

**Part 1**

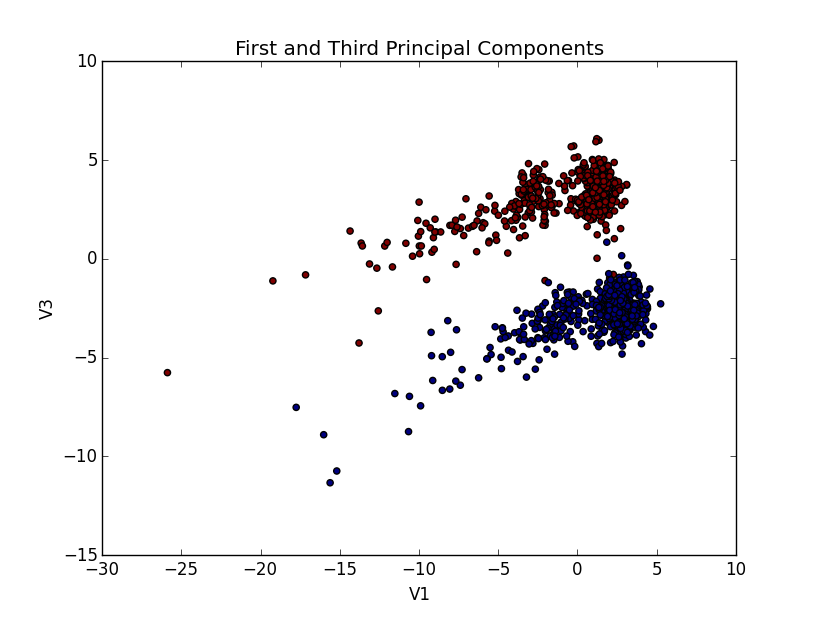
(a). No deliverable

(b). Plot included



(c). The fact that the clusters are fairly close to each other indicate that “V1” doesn’t do a particularly good job of finding a differentiating characteristic of the data. Instead “V2” is capturing a lot of the variance in these population clusters. In addition, the orange and green clusters are not strongly influenced by the properties of the data captured in the first principal component. It seems that the dataset is separating those individuals of pure African ancestry from those of African ancestry from the Carribeans and the United States.

(d). Plot included



(e). The third principal component differentiates the gender of individuals in all the populations.

(f). [737 445 635] Assuming a strong correlation among these three largest indices (and recognizing that the 3rd principal component encodes a signal for gender), a deviation in the mode of these indices should strongly correlate to a given gender (i.e. a ‘1’ might indicate male).

(g). For a given column (feature) in a specific individual (row), we are essentially interested in giving higher importance to a feature that deviates strongly from the mode, as this is likely to be a feature that distinguishes our data and we want to pay especially close attention to these differentiating features. Thus, we can compute a per column frequency of the 4 nucleobases, (i.e. A: 10, B: 8, C: 4, D: 3). Then the column for a given individual in our Y matrix will indicate the distance from the mode. In the example given, a value of A in a column will indicate 0, B will indicate 1, C will indicate 2, and D will indicate 3.

**Part 2**