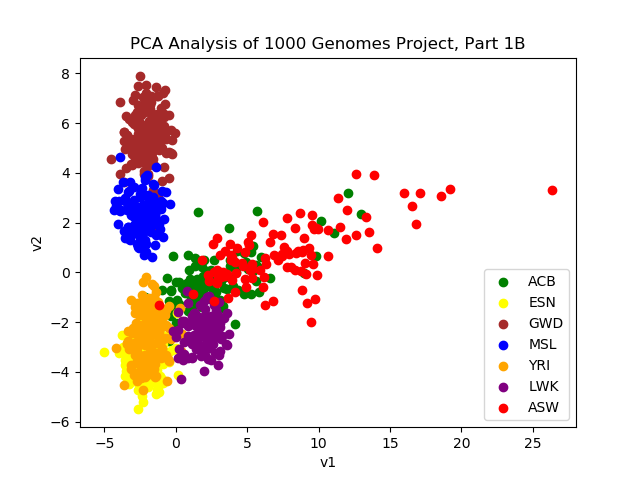
CS168 Project 4

Henry Lin, Kaylee Bement

1. A.

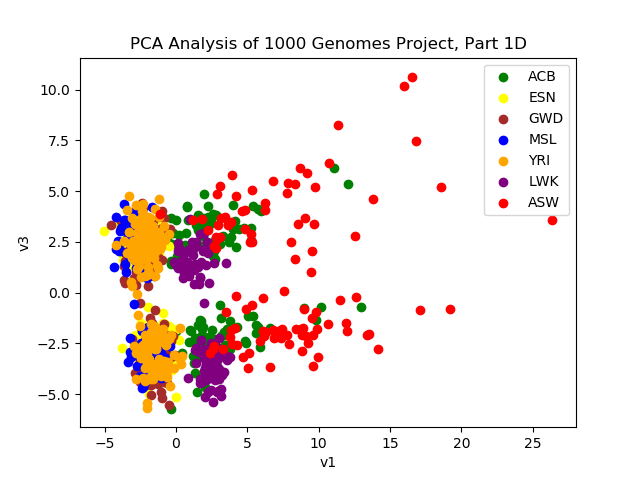
B.

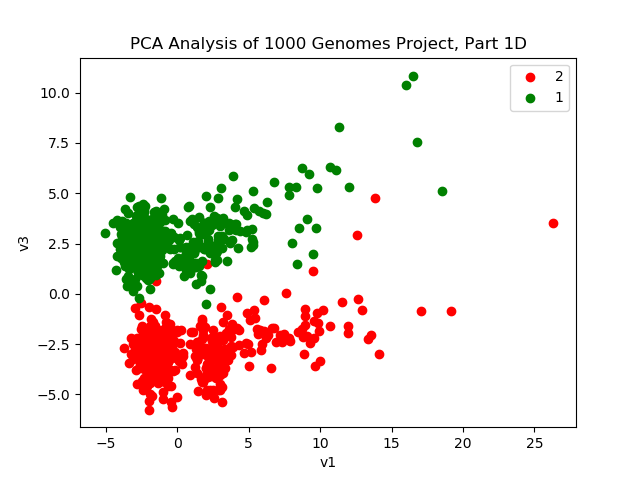


C.

From the plot, it appears that the first two components capture the geographic proximity of populations to each other coupled along with rate of mixture between different groups. For instance, it is likely that MSL and GWD are both populations that are geographically separated from the others, leading to the separate clusters with little or no overlap. On the other hand, YRI and ESN could plausibly be populations that are located close geographically and have experienced a lot of inter-population reproduction. For ASW, it is likely that it is a more nomadic population due to the decentralized cluster that it forms.

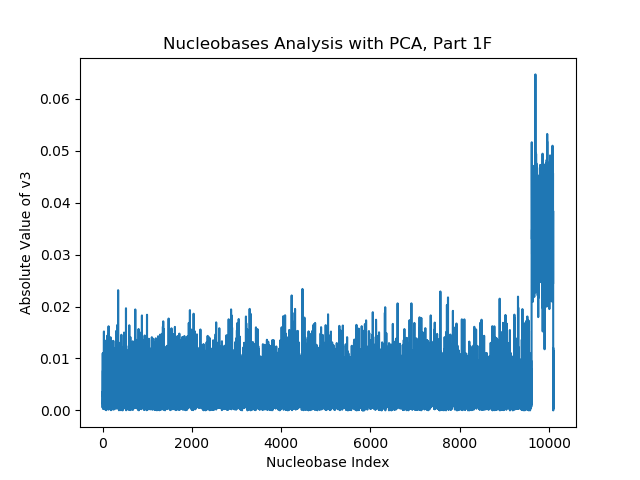
D.





For the second plot, the 1s represent male and 2s represent female.

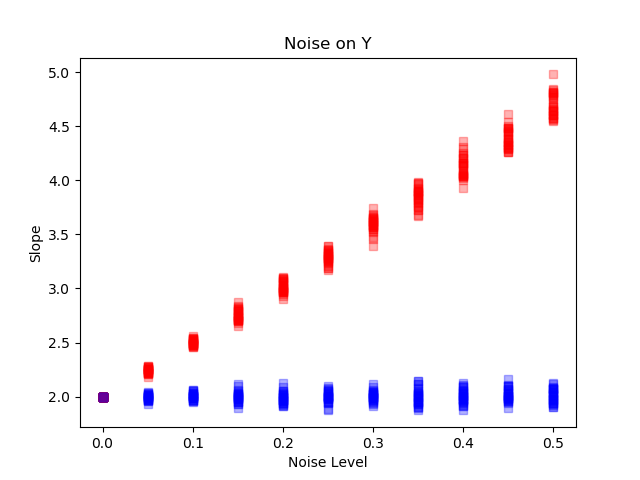
E. V3 captures the sex of the individual.

F.

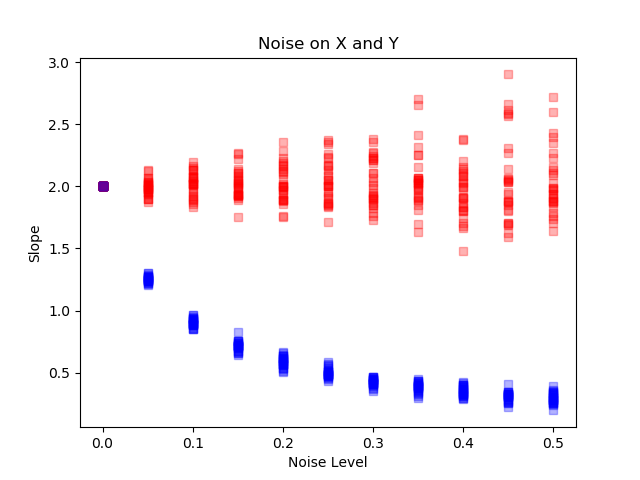
There is a large spike near the end of the nucleobase index. This is most likely because the sex of the individual is determined by the two sex chromosomes, which can be XX or XY. Since there are located at the end of the nucleobase index, it makes sense that there is a spike at the end since they are the determining factors for this component.

1. B. //

C.



D.



E.

In PCA, all coordinates play a symmetric role in computing the best fit. Thus, PCA does poorly with noise in only Y because it has to fit itself equally between a non-noisy X and a noisy Y, while it does well with noise in X and Y because the noise affects both coordinates approximately equally, making it easier to find the pattern.

LS does poorly with noise in X and Y because ??????