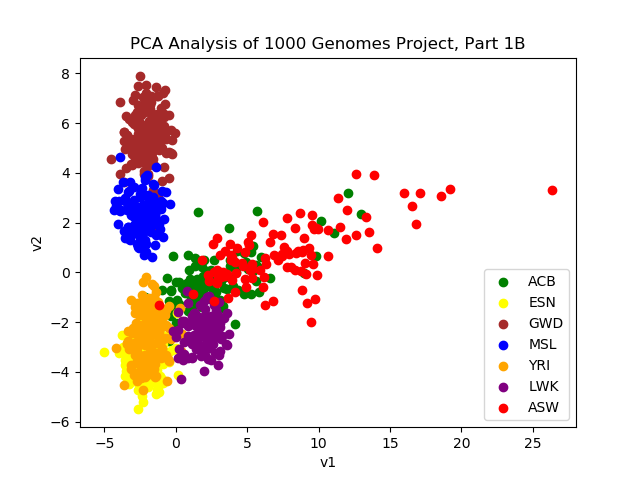
CS168 Project 4

Henry Lin, Kaylee Bement

1. A. Warmup

B.



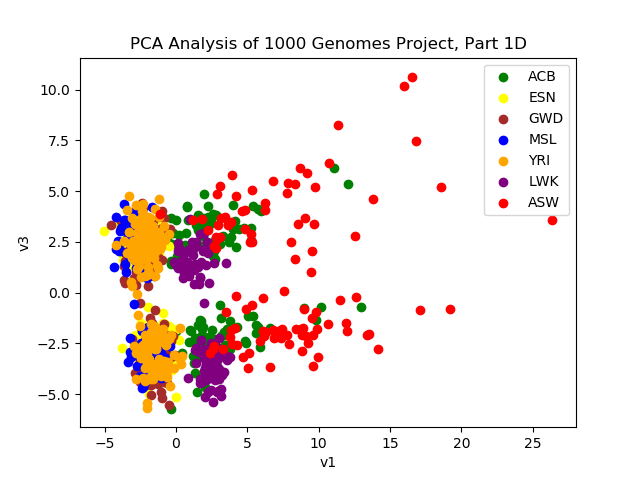
C.

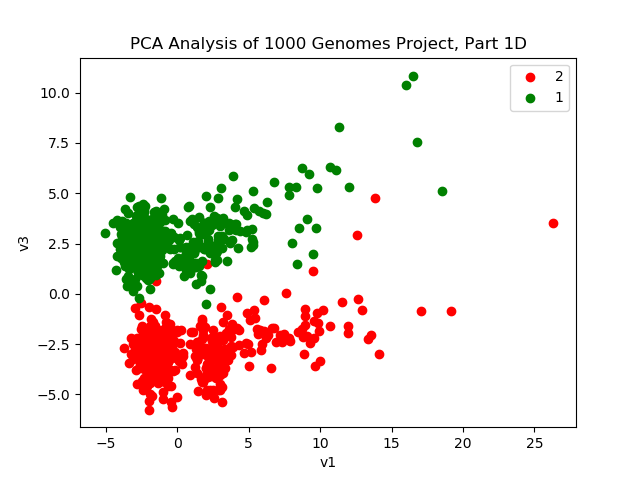
From the plot, it appears that the first two components capture the geographic proximity of populations to each other coupled along with rate of mixture between different groups. For instance, it is likely that MSL and GWD are both populations that are geographically separated from the others, leading 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.

With these hypotheses, we can examine a map of Africa and what populations each cluster corresponds to verify our conjectures. LWK and GWD are both groups located in far West Africa (Gambia and Sierra Leona) that are geographically separated from the other populations but close to one another, explaining their distinct genomes. ESN and YRI are both groups located in Nigeria, and their geographic proximity to each other explains the strong overlap. ACB and ASW cover the Barbados and African Americans, and this explains the more spread out cluster because these groups have geographically moved away from the others across time and possible intermingling with populations not present may have caused the distinct genomes. The African American genome spread could be attributed to the melting pot that is the United States. The ACB and ASW clusters’ proximity to each other cluster also suggests which groups they may have descended from.

It is hard to define a specific characteristic each component captures such as longitude or latitude from lecture, but something like those two characters can be seen when plotting the 2 components above.

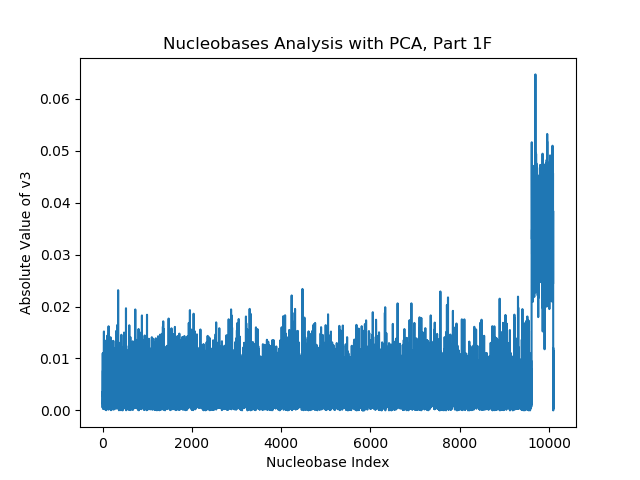
D.





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

E. V3 captures the sex of the individual.

F.

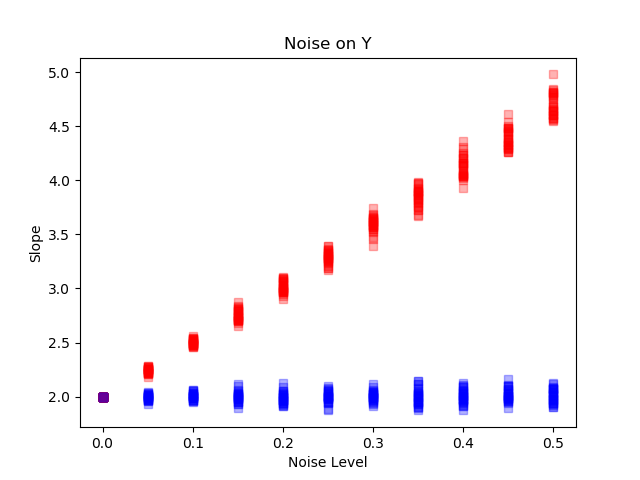
There is a large spike near the end of the nucleobase index. This is most likely because the sex of the individual is determined by the two sex chromosomes, which can be XX or XY. Since there are located at the end of the nucleobase index, it makes sense that there is a spike at the end since they are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

1. B.

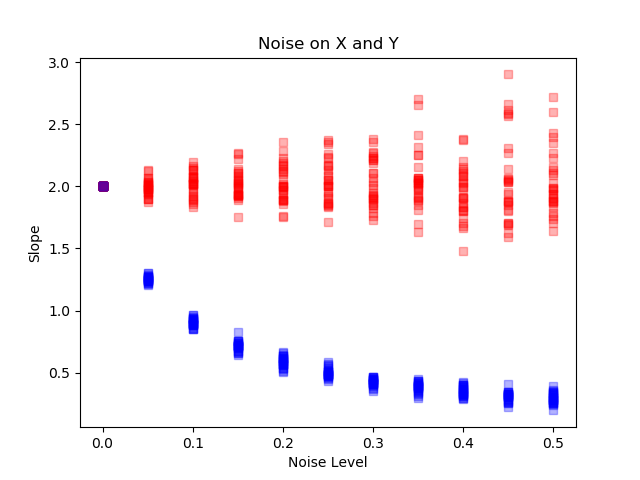
Least squares would recover the line y = 0.5 because in the defined uniform distribution, the average distance from the best fit line will be 0.25 when y= 0.5.

PCA would recover the line y = x because it recovers the direction in which the data has the maximum variance, which would be the diagonal of the square.

C.



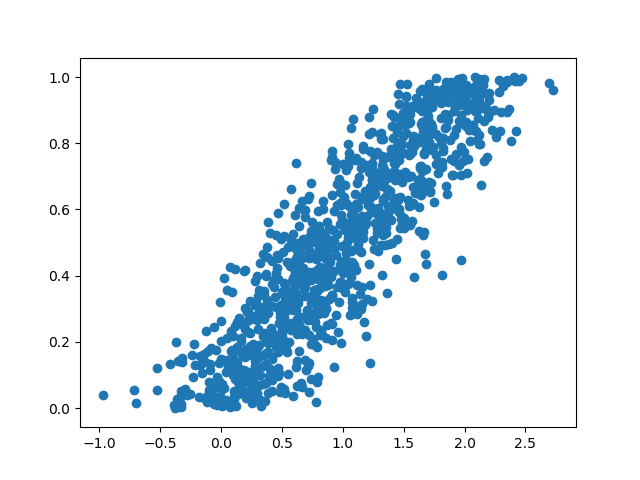
D.



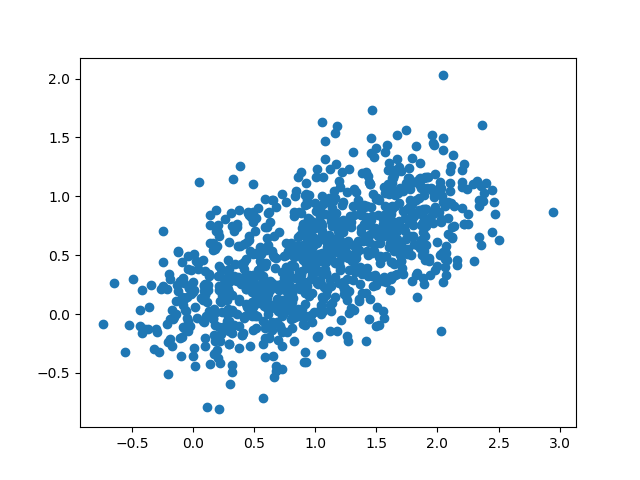
E.

Least Squares does really well with noise on just Y because it only cares about the vertical distance between the best fit lien and the point. If each point on line receives some noise from a normal distribution, the average overall all the points will still be the line y = 2x because some points will be shifted up and other will be shifted down, canceling each other out.

PCA does not perform well with noise on just Y because the direction of max variance changes when there is just noise on Y. With the points solely up and down, we can imagine the noisy points scattered in a parallelogram area around the original as shown below. The long diagonal of this parallelogram is now the direction of maximum variance, so it is what PCA will recover. As one can imagine, as noise increases, the more deviated this long diagonal line becomes from the true y = 2x, causing PCA to perform badly.



PCA does well when there is noise on X and Y because the parallelogram above does not exist. Instead, the noisy data points from a oval-ish shape around the true line, and since the direction of maximum variance for an oval shape is its longer axis, it matches up with the true line. Look at the below graph when the noise level c = 0.1



Least Squares starts to fail as the noise level increases because the oval starts turning into a more circle-like shape around the true line. The is problematic because least square regression will actually fit a horizontal line to a perfect circle to keep the maximum distance equal to the radius. This leads to the slope it returns decreasing as the noise increases, which is evident in the plot from 2d. This can be seen when you compare the above plot with c = 0.1 to the below plot with c = 0.4

