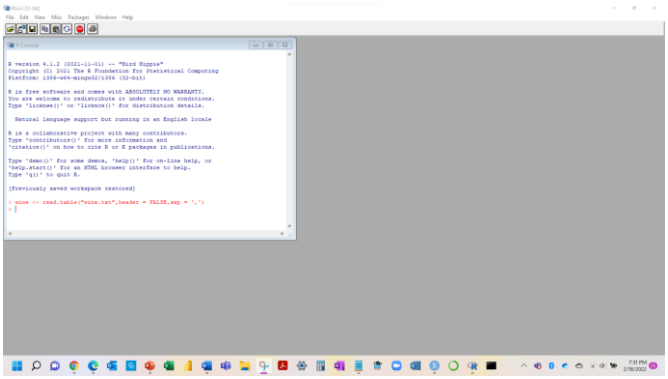
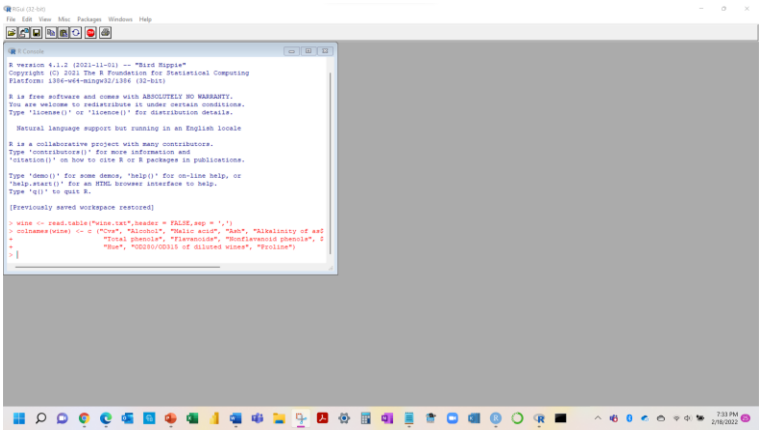
Josh Cubero
Principal Components Analysis
MSBA 324
Golden Gate University

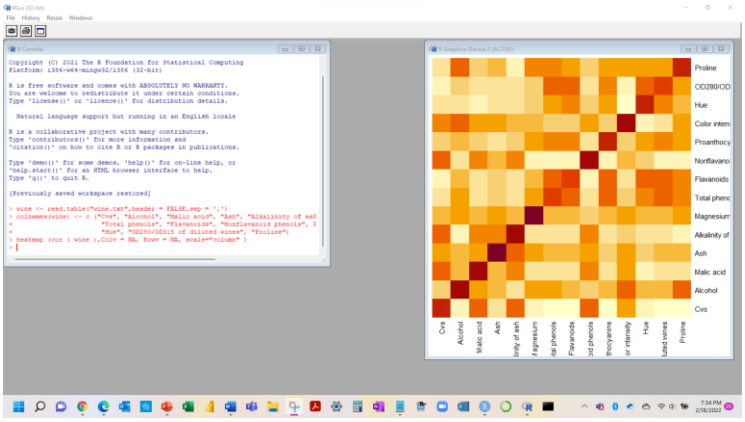
1. This image depicts load the wine dataset to the workspace, note I changed directory to the folder where the file is located to keep the syntax shorter.



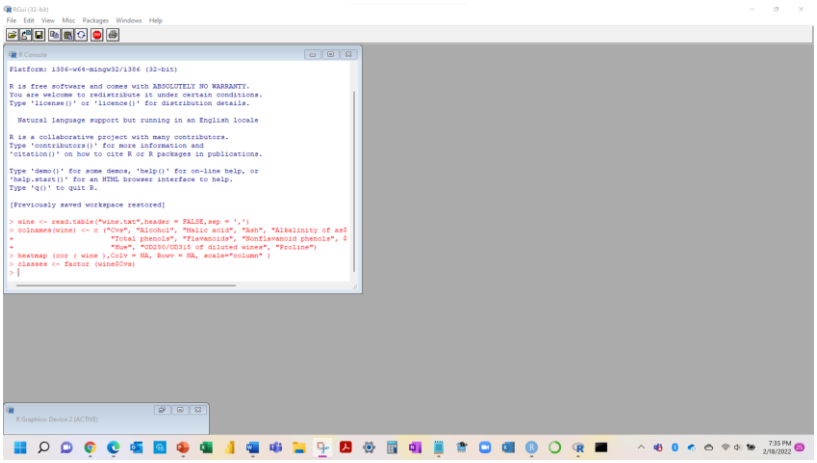
2. This image assigns column names to the wine dataset, the wine did not have column names in the txt file.



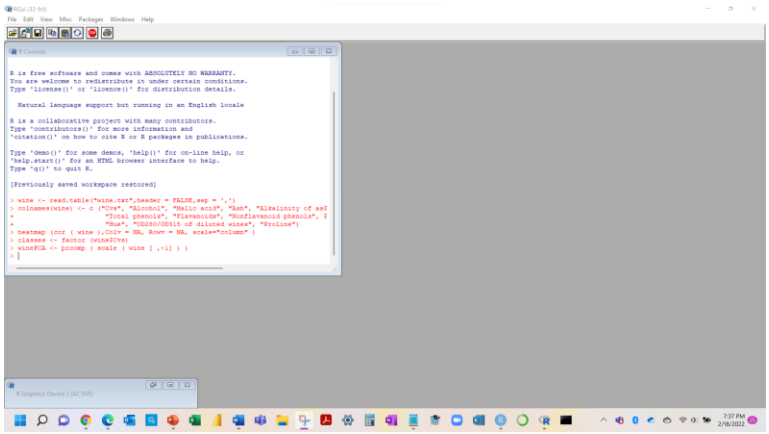
3. The heat map depicted below is a visual representation of Excel's correlation matrix. The intent of the heat map is the display the correlation between variables in the dataset. Identifying correlations is important because strongly correlated explanatory variables cause collinearity in a regression analysis. Collinearity can cause R and R squared in a regression to be inflated and introduce erroneous predictions. The color intensity in the heatmap displays the strength of the correlation. Strongly correlated variables will have a darker coloration. You'll notice that the center diagonal squares are the most intense because that is the cross section between a given variable. Any squares darker than the lightest color should be cause for concern and be further investigated.



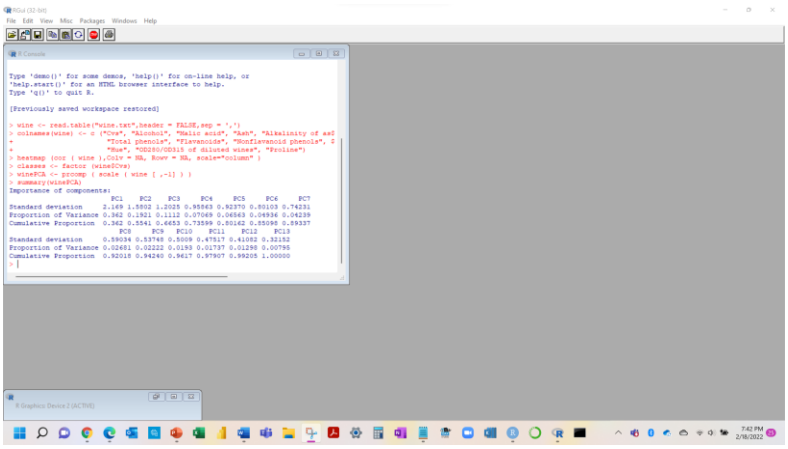
4. The code depicted below assigns classes to the 3 cultivars via the factor command.



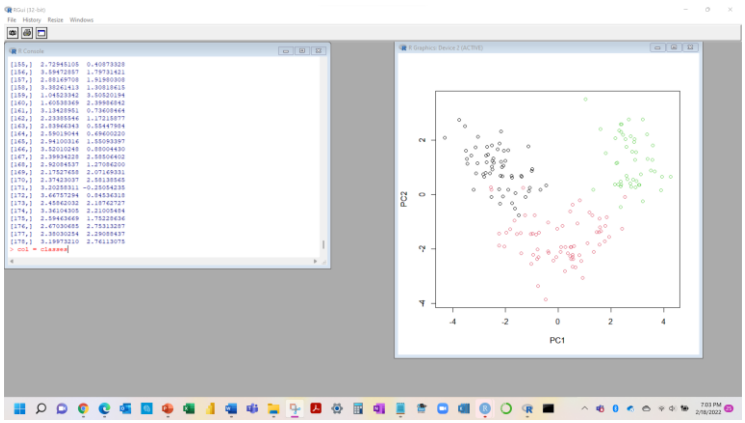
5. This snippet depicts normalizing the wine dataset with the scale function. The process normalizes the data so that the mean is 0 and standard deviation is 1.



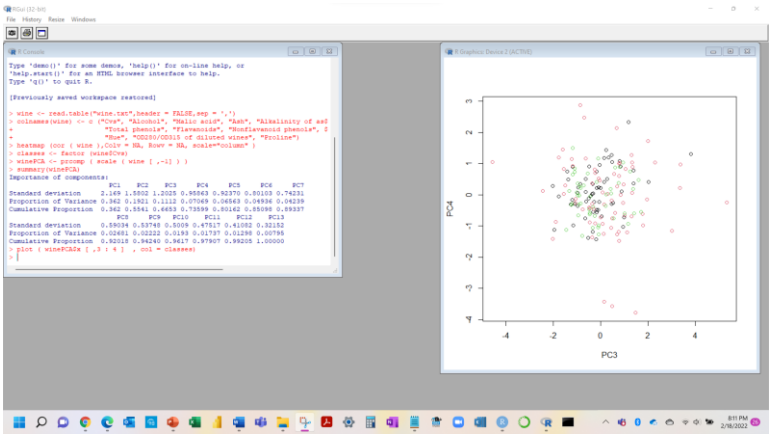
6. The following describes number of PC variables required to explain a specific amount of variance:
    a. 50% requires 2 variables, as we can see, the cumulative proportion of PC2 is 55.41%.
    b. 80% requires 5 variables, a look at PC 5 indicates the cumulative proportion is 80.01%.
    c. 90% requires 8 variables as witnessed by the cumulative proportion of PC8 at 92.018



7. PC1 vs PC2 with class variables as colors.



8. Plot of PC3 vs PC4.



9. One look at PC1 vs PC2 and PC3 vs PC4, one would immediately see that PC3/PC4 plots are nearly identical. An observer might say that there's very little variation between the two plots, and they'd be correct. Comparing the PC1 vs PC2, we see very distinct variation between the values of the plots. Thus, more variation in the dataset is explain by analyzing PC1 vs PC2 than that of PC3 vs PC4.