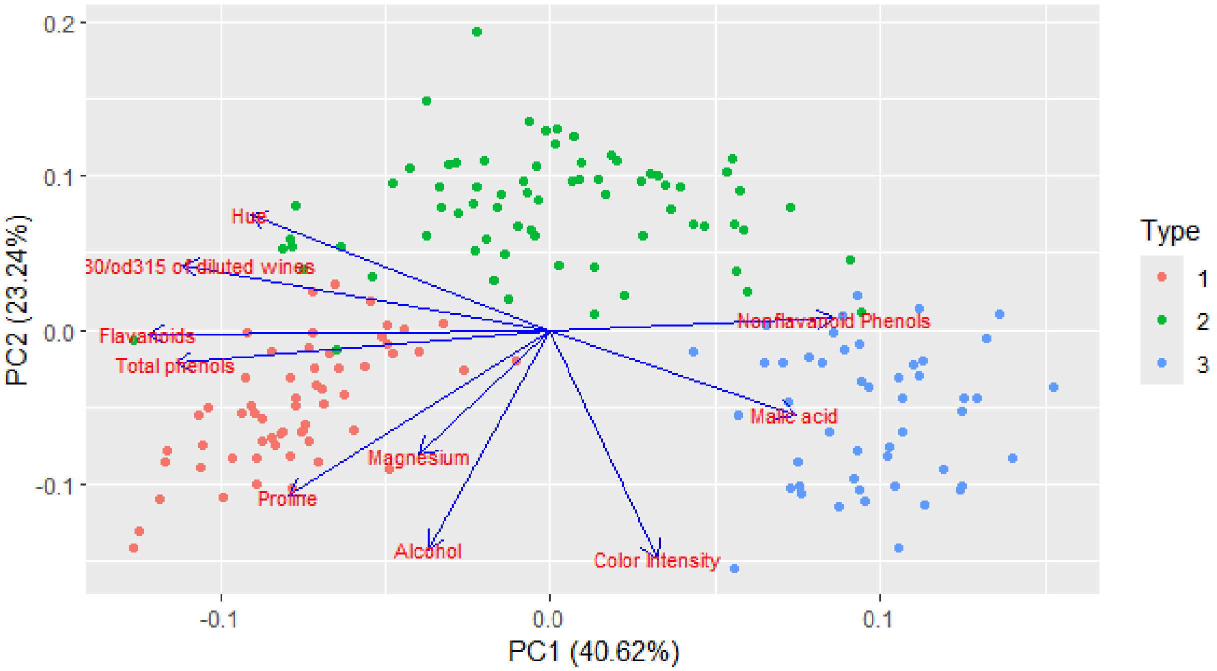
Mike Dowd
lab04

1. Plot of dataset and 1st and 2nd principal components



2. It seems Flavanoids, Total Phenols, and Od280/Od315 of diluted wines influence PC1 the most, all with correlations of around -0.4.

```
          Alcohol                    Malic acid
        -0.1387659                      0.2774631
         Magnesium                   Total phenols
         -0.1482252                    -0.4234687
         Flavanoids           Nonflavanoid Phenols
         -0.4555529                     0.3215373
         Color Intensity                 Hue
         0.1214953                    -0.3392301
Od280/od315 of diluted wines             Proline
         -0.4166142                    -0.2968863
```

3. Model results using all 13 attributes:

```
              actual
predicted  1   2   3
        1 55   5   3
        2  0  49  10
        3  4 17 35

     recall precision        f1
1 0.8730159 0.9322034 0.9016393
2 0.8305085 0.6901408 0.7538462
3 0.6250000 0.7291667 0.6730769
[1] "accuracy =  0.780898876404494"
```

4. Model results using first 3 PCs:

    ```
            actual
    predicted  1  2  3
            1 59  2  0
            2  0 67  1
            3  0  2 47

         recall precision        f1
    1 0.9672131 1.0000000 0.9833333
    2 0.9852941 0.9436620 0.9640288
    3 0.9591837 0.9791667 0.9690722
    [1] "accuracy =  0.971910112359551"
    ```

    Notably much better than the original model, probably from normalization and standardization(since I’m using KNN which is based on distance).

5. Alcohol, Magnesium, and Color Intensity all had extremely small correlations in the rotation matrix for the first PC, with their magnitudes being around 0.1

Model results with these variables removed and rerunning PCA(model created using first 3 PCS):

```
        actual
predicted  1  2  3
        1 57  6  0
        2  2 62  0
        3  0  3 48

     recall precision        f1
1 0.9047619 0.9661017 0.9344262
2 0.9687500 0.8732394 0.9185185
3 0.9411765 1.0000000 0.9696970
[1] "accuracy =  0.938202247191011"
```

As we can see, the results are a little worse, but we removed a lot of data and still got very good accuracy, which could be useful to speed up the process if we have large datasets.

Of the three KNN models, the ones utilizing the first 3 PCs with standardized data performed significantly better than the raw data. However, the model utilizing all 13 features performs a notable margin better than the one with features of low importance in the first PC removed.