My Name Biao Feng(myNetID: biaof2)

IE598 MLF F18

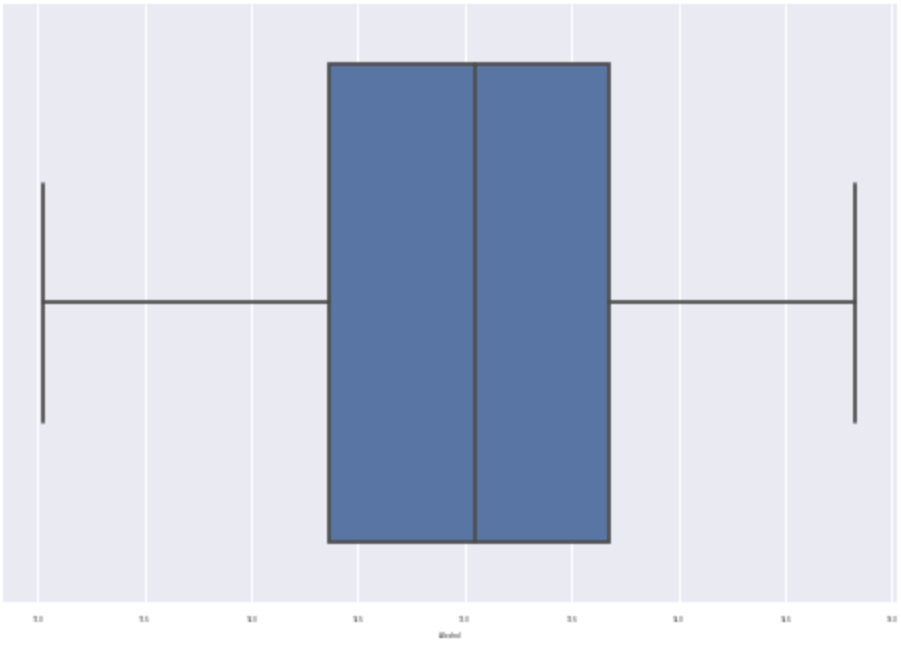
Module 5 Homework (Dimensionality Reduction)

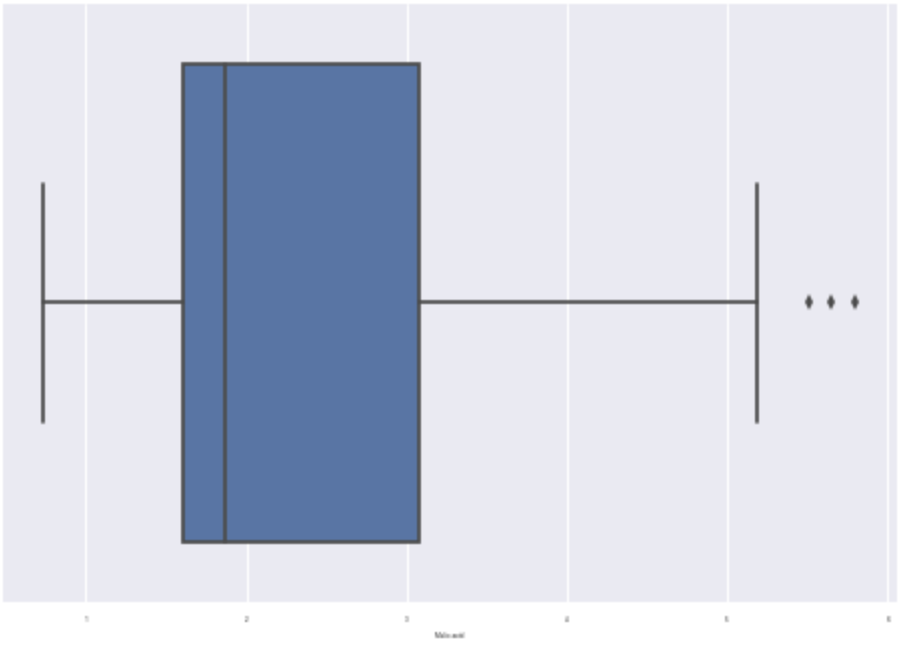
Part 1: Exploratory Data Analysis

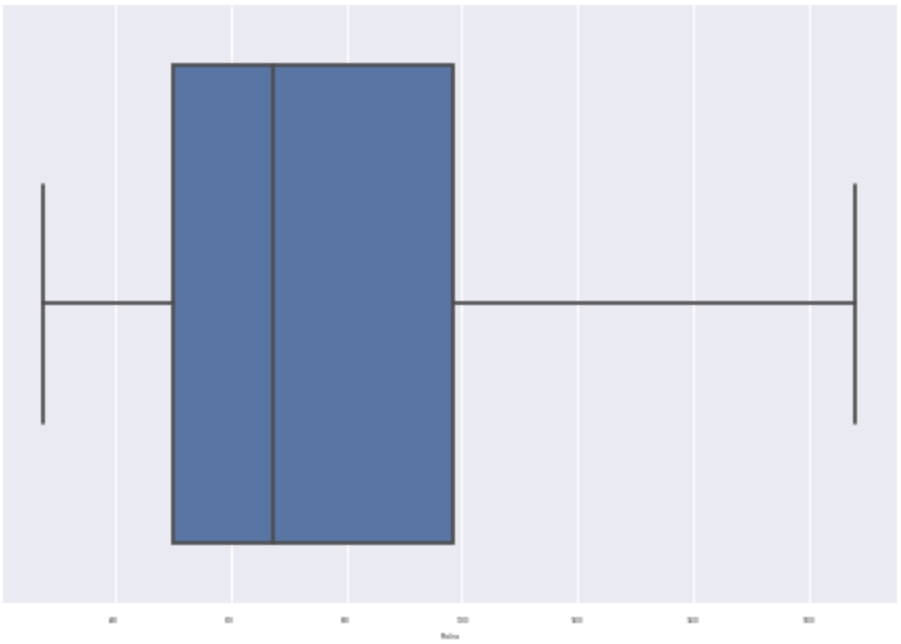
Describe the data sets sufficiently using the methods and visualizations that we used previously. Include any output, graphs, tables, heatmaps, box plots, etc. that you think is necessary to represent the data. Label your figures and axes. DO NOT INCLUDE CODE, only output figures!

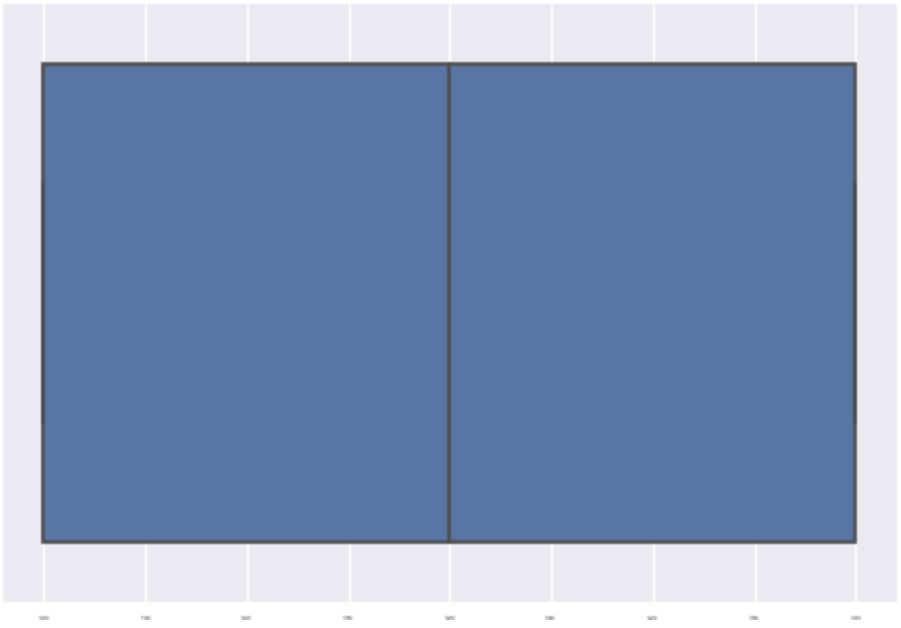
Split data into training and test sets. Use random_state = 42. Use 80% of the data for the training set. Use the same split for all experiments.

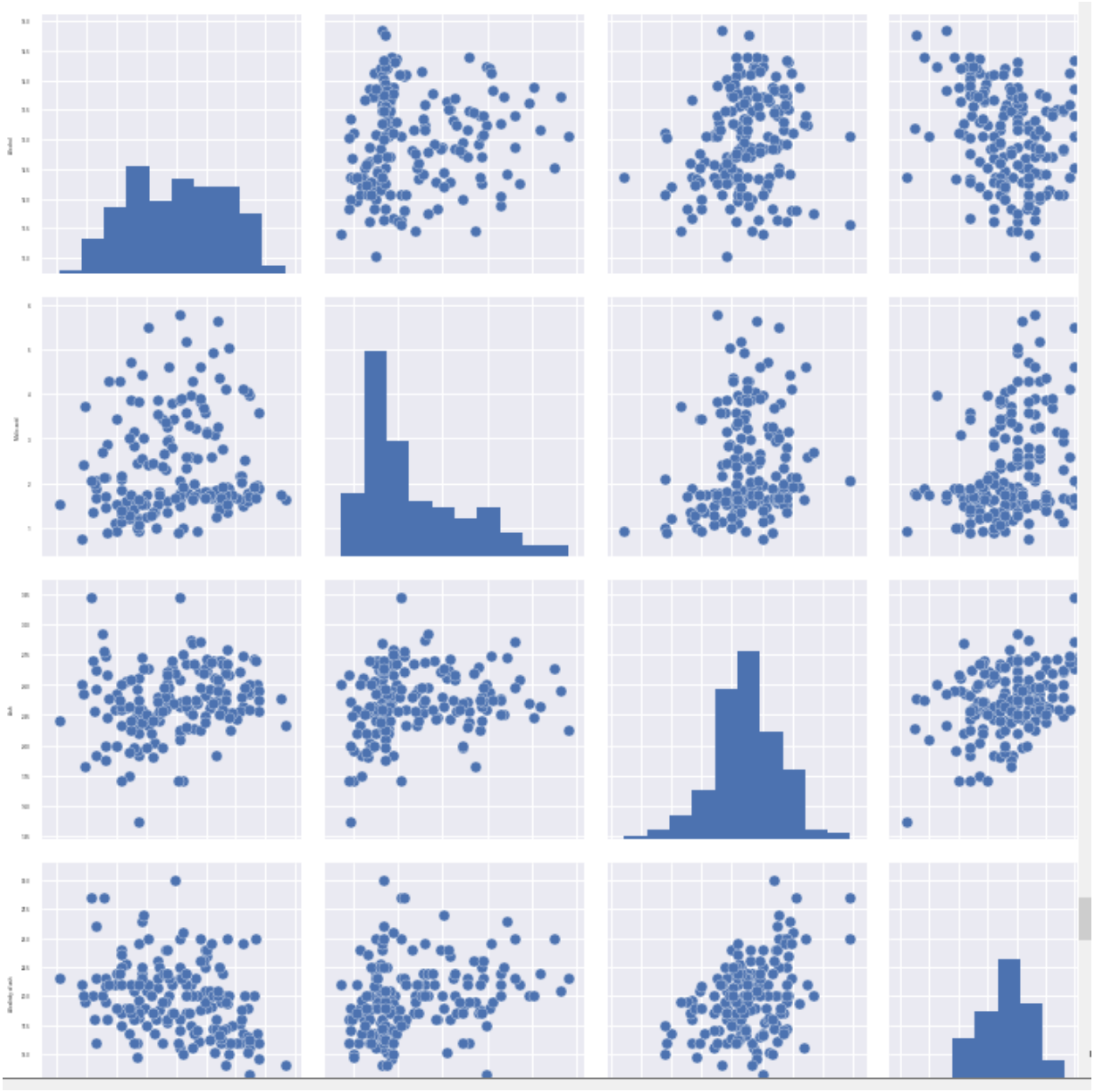
```
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
Alcohol                         178 non-null float64
Malic acid                      178 non-null float64
Ash                             178 non-null float64
Alcalinity of ash               178 non-null float64
Magnesium                       178 non-null int64
Total phenols                   178 non-null float64
Flavanoids                      178 non-null float64
Nonflavanoid phenols            178 non-null float64
Proanthocyanins                 178 non-null float64
Color intensity                 178 non-null float64
Hue                             178 non-null float64
OD280/OD315 of diluted wines    178 non-null float64
Proline                         178 non-null int64
Class                           178 non-null int64
dtypes: float64(11), int64(3)
memory usage: 19.5 KB
```

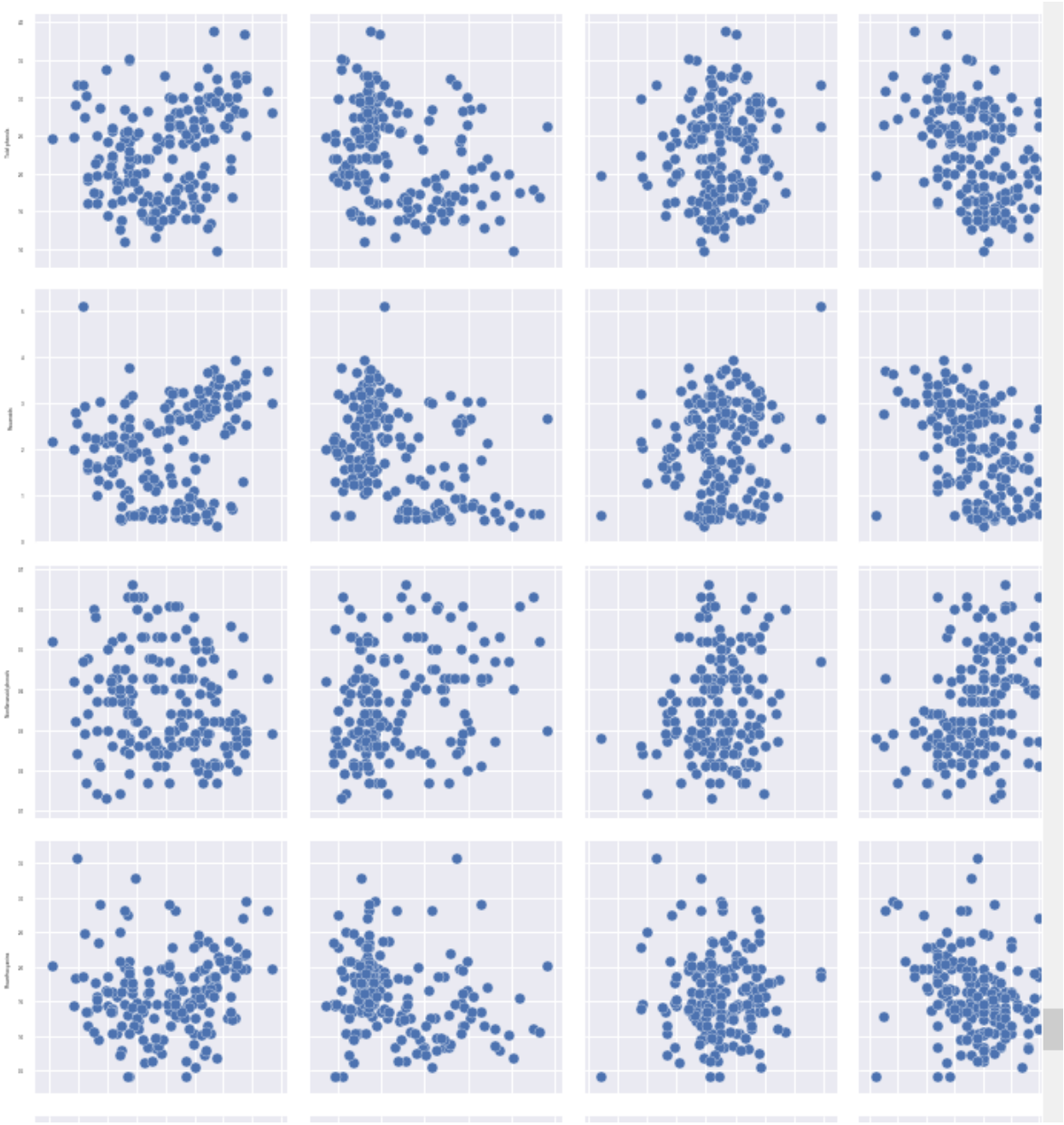


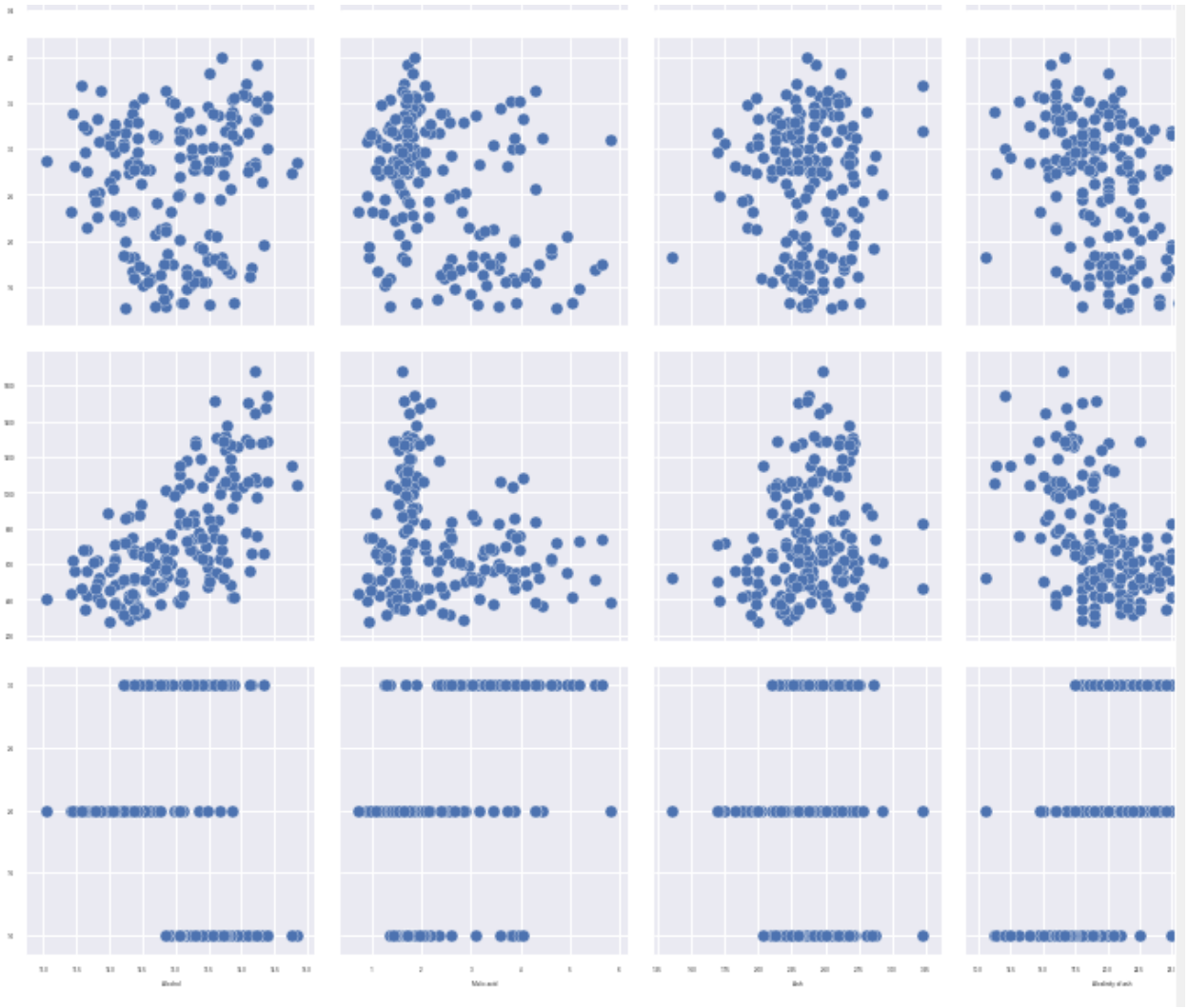


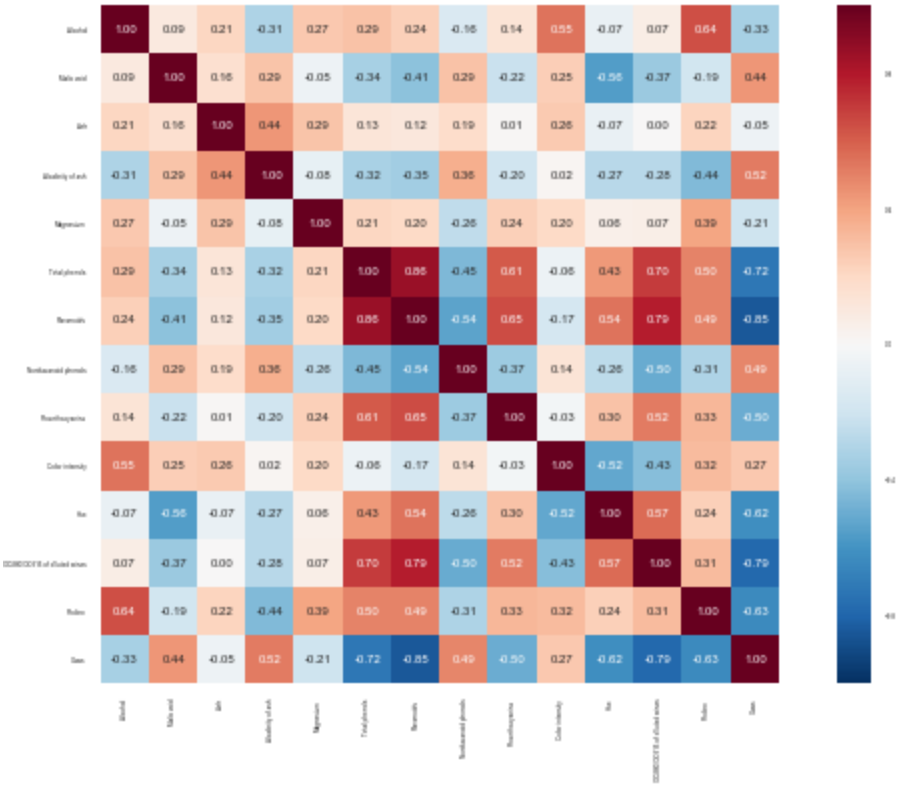


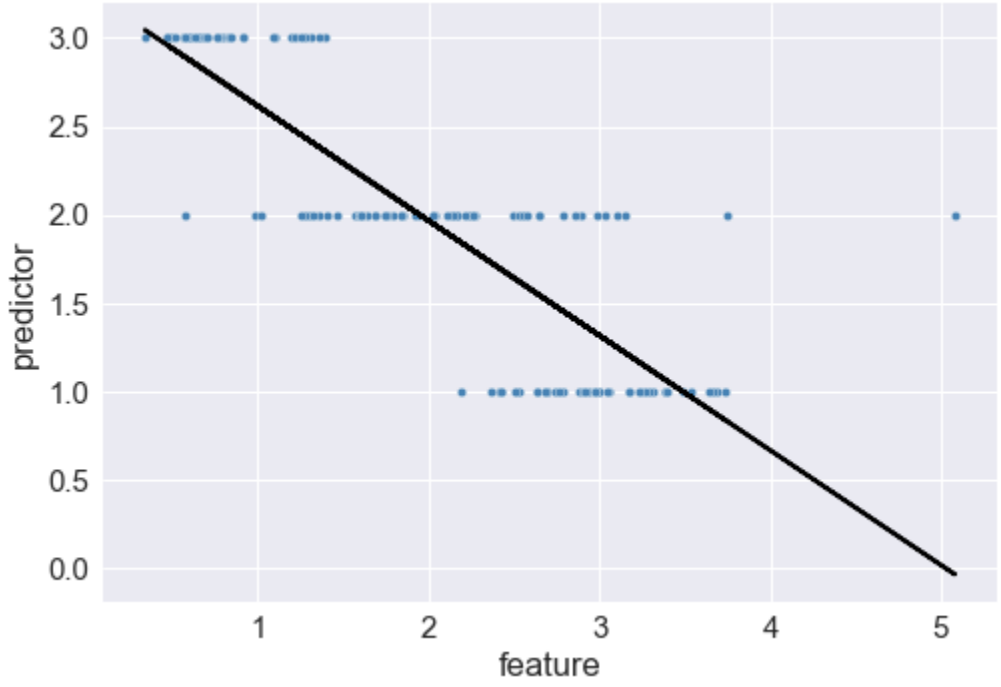












Part 2: Logistic regression classifier v. SVM classifier - baseline

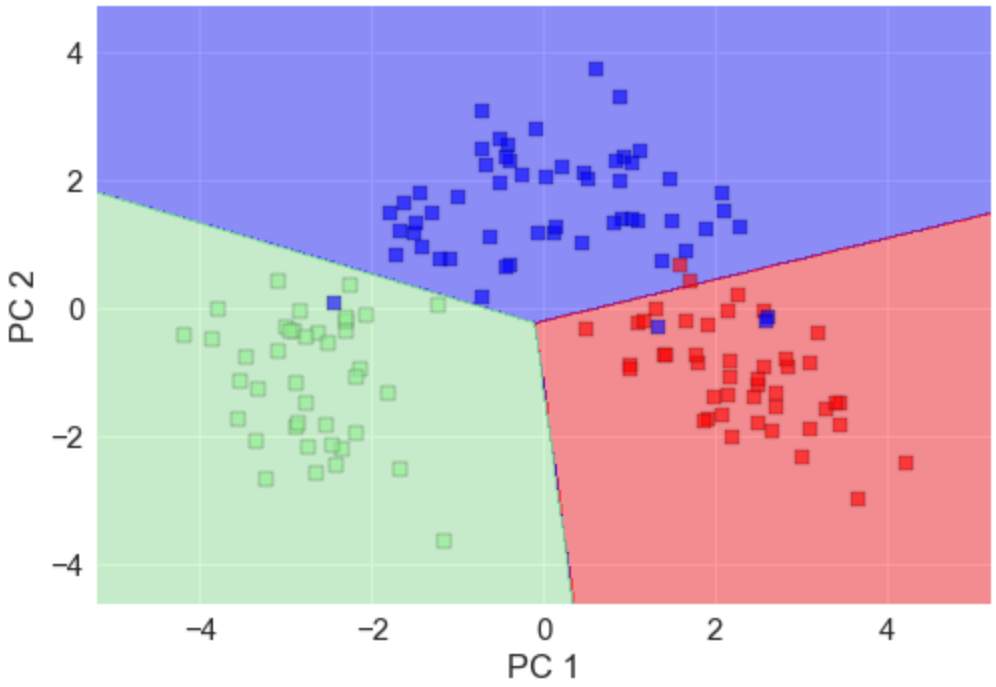
Fit a logistic classifier model to both datasets using SKlearn. Calculate its accuracy score for both in sample and out of sample (train and test sets). (You may use CV accuracy score if you wish).

Fit a SVM classifier model to both datasets using SKlearn. Calculate its accuracy score for both in sample and out of sample (train and test sets). (You may use CV accuracy score if you wish).

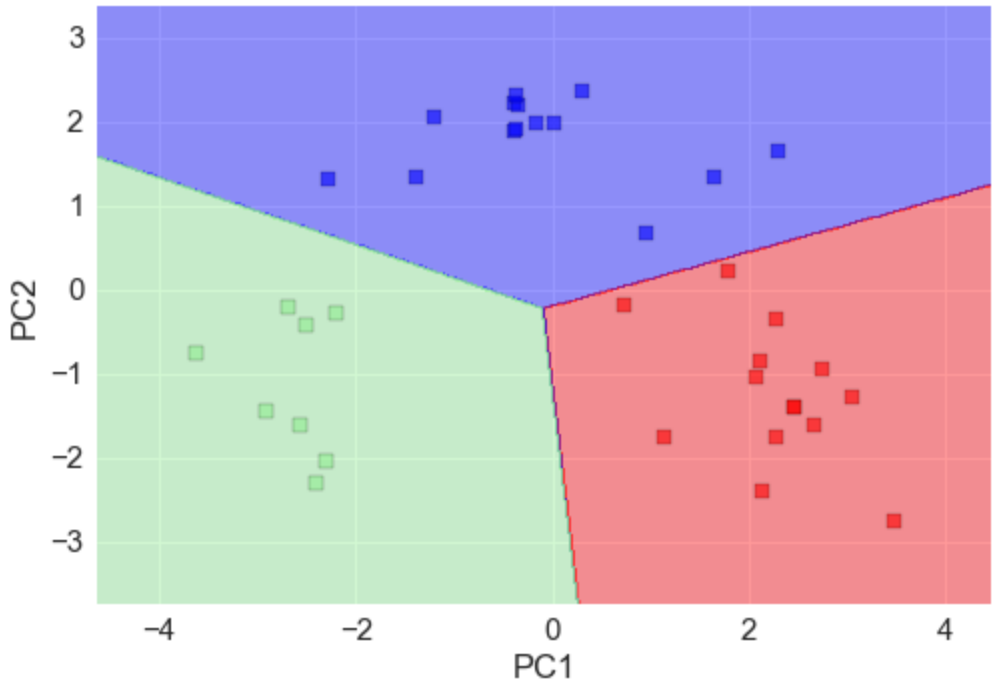
```
sv.score(X_test_std, y_test))
lr train R^2: 1.00
lr test R^2: 1.00
sv train R^2: 0.99
sv test R^2: 0.97
```

Part 3: Perform a PCA on both datasets

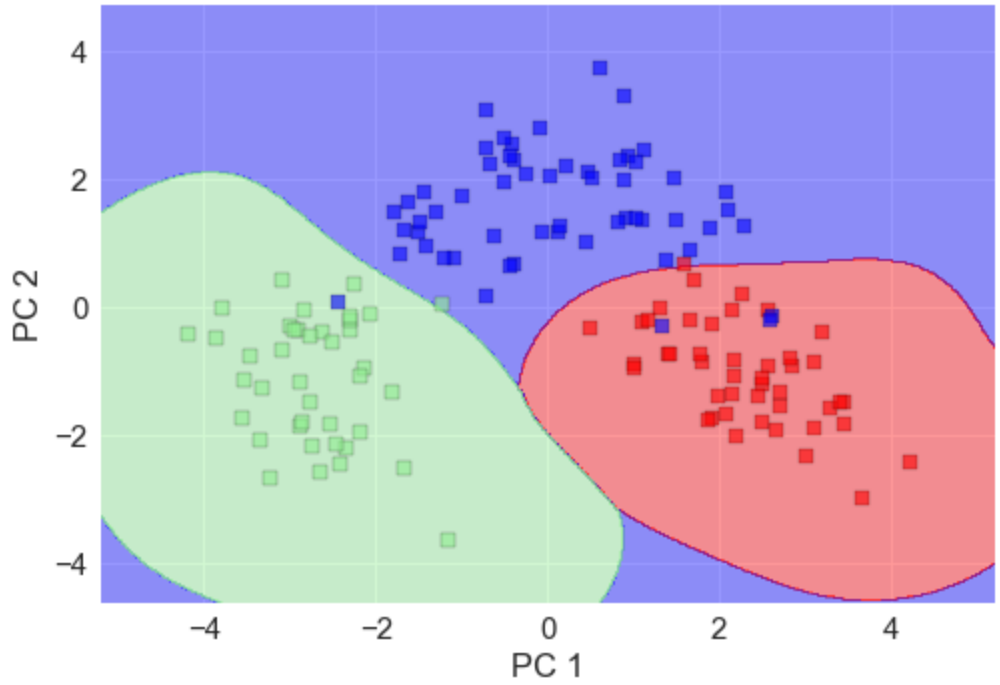
Refit both a logistic and SVM classifier on the PCA transformed datasets. You may choose to use only 2 components, or select a higher appropriate intrinsic dimension. Calculate accuracy scores for both in sample and out of sample (train and test sets) on both datasets.

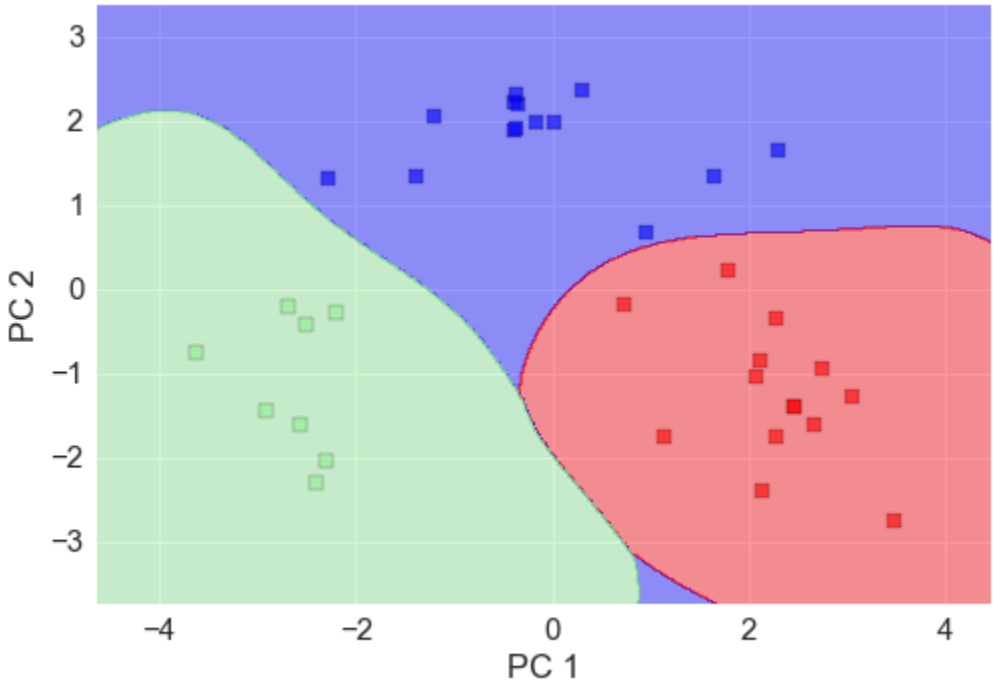



```
pca lr train R^2: 0.96
```



```
pca lr test R^2: 1.00
```

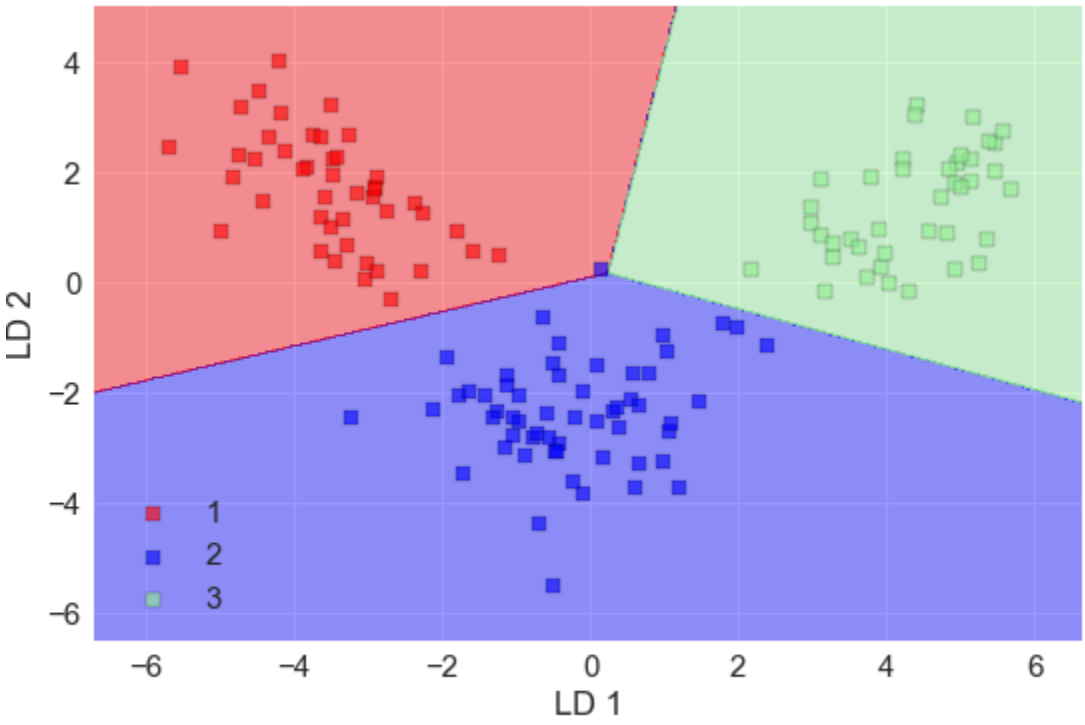




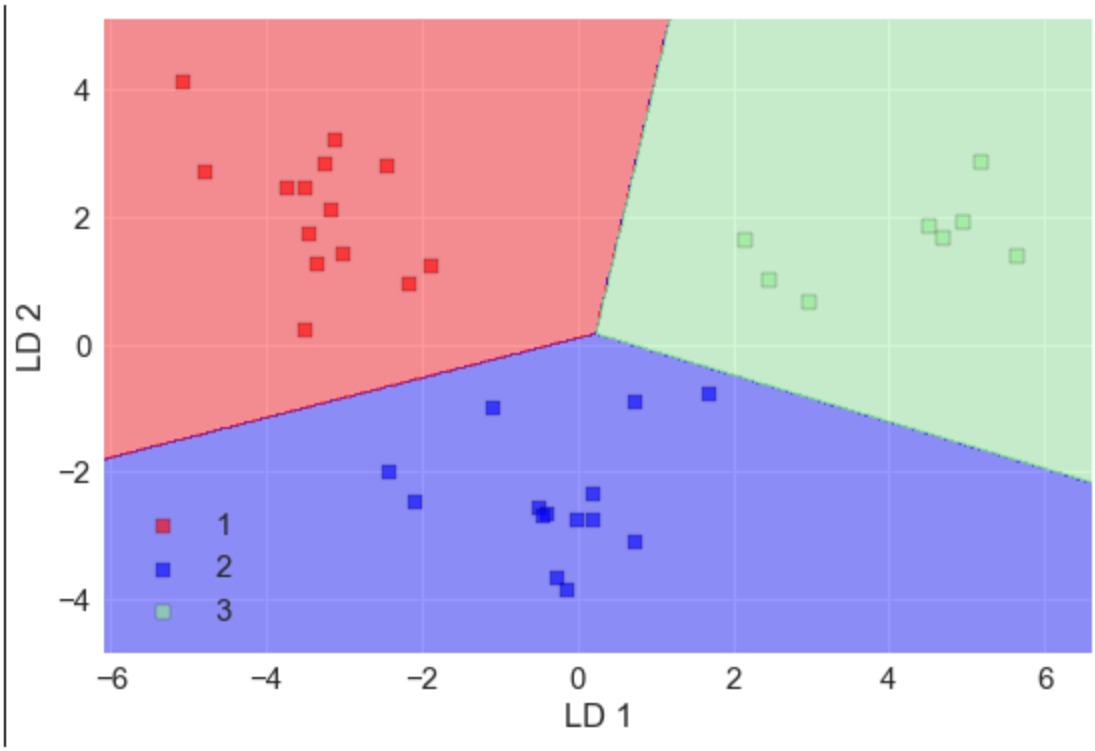

```
pca sv test R^2: 1.00
```

Part 4: Perform and LDA on both datasets

Refit both a logistic and SVM classifier on the LDA transformed datasets. You may choose to use only 2 discriminants, or select a higher appropriate number. Calculate accuracy scores for both in sample and out of sample (train and test sets) on both datasets.

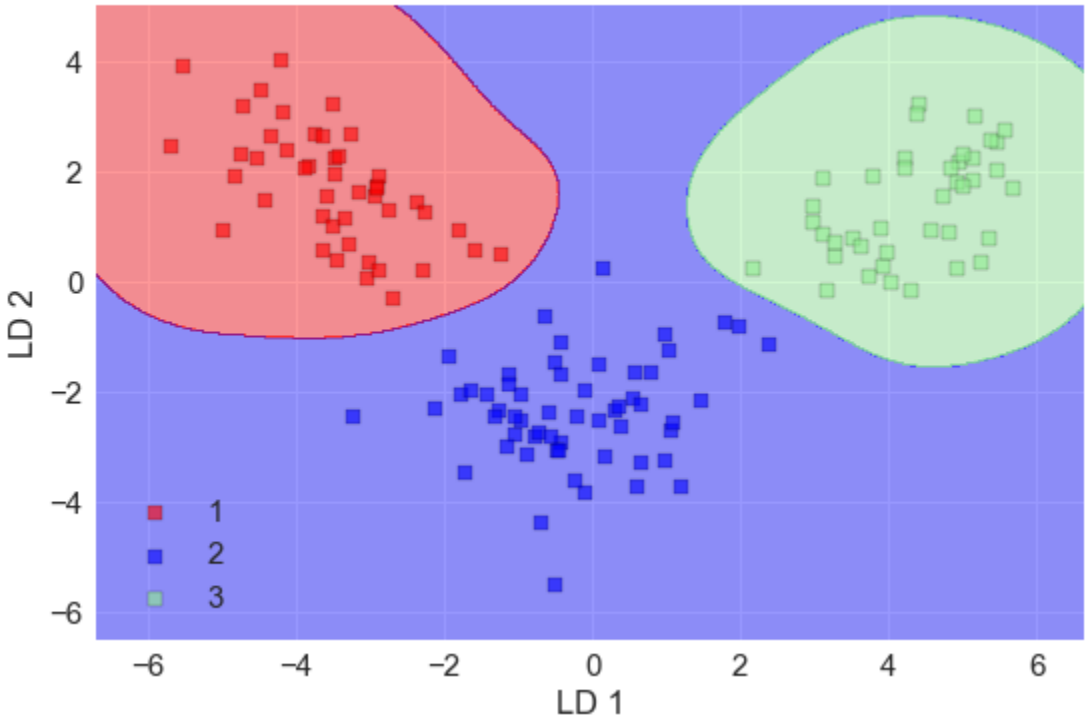


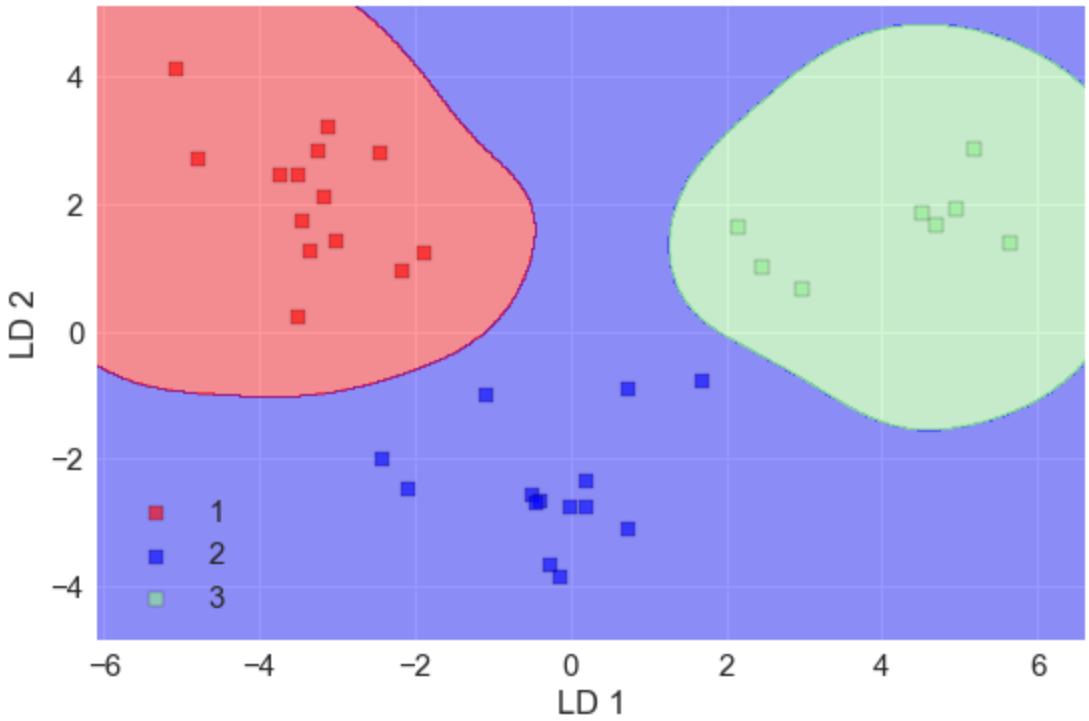

```
lda lr trainR^2: 0.99
```




```
lda lr test R^2: 1.00
```


```
lda lr test R^2: 1.00
```

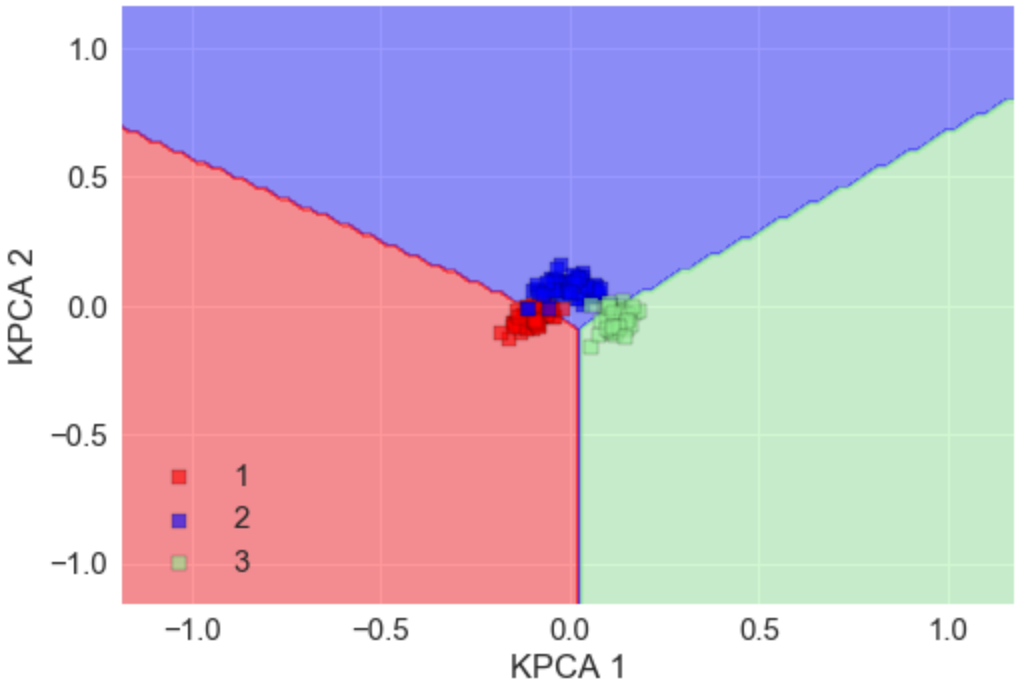



```
lda sv test R^2: 1.00
```

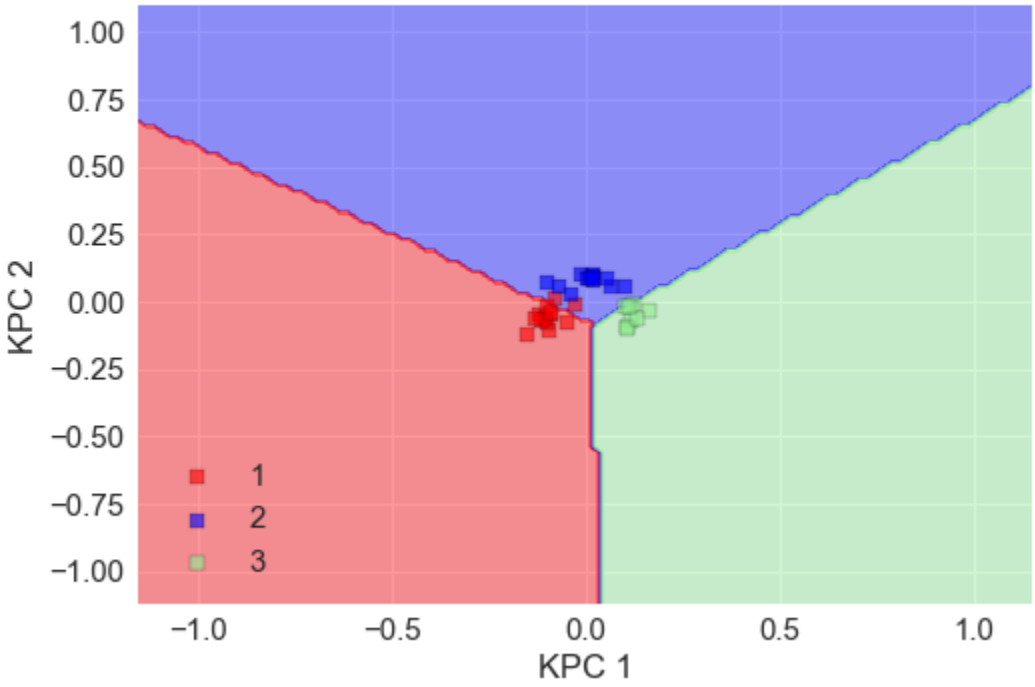
Part 5: Perform a kPCA on both datasets

Refit both a logistic and SVM classifier on the kPCA transformed datasets. Use the rbf kernel. Test several different values for Gamma. Calculate accuracy scores for both in sample and out of sample (train and test sets) on both datasets.

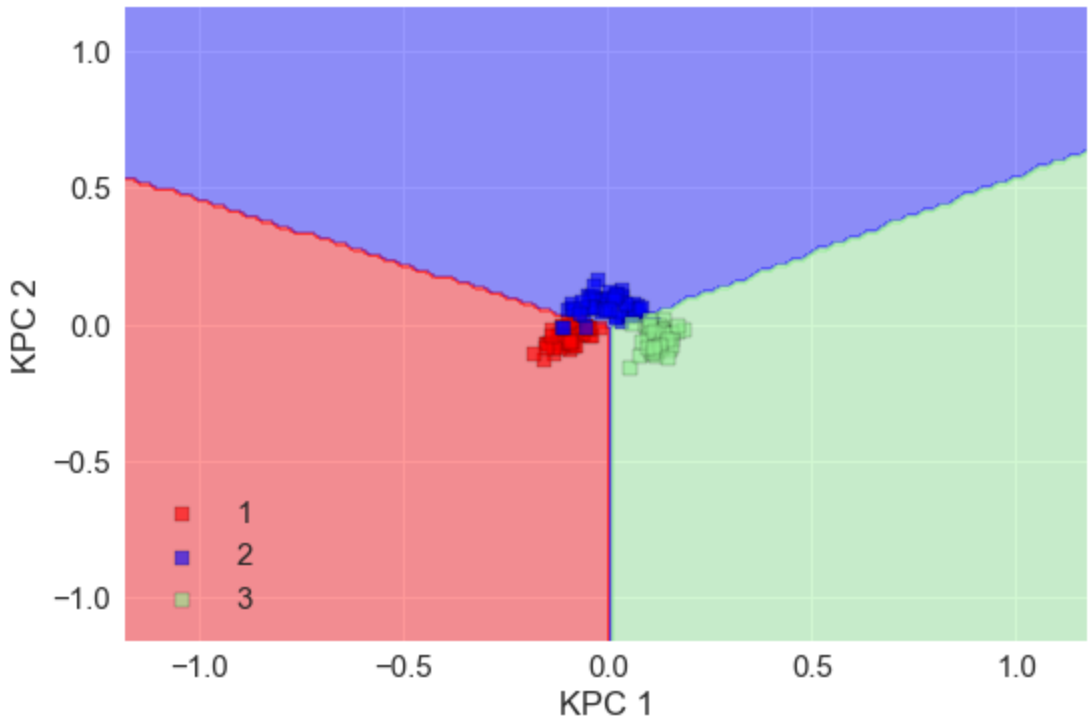
gamma 0.001



kpca lr trainR^2: 0.84
gamma 0.001

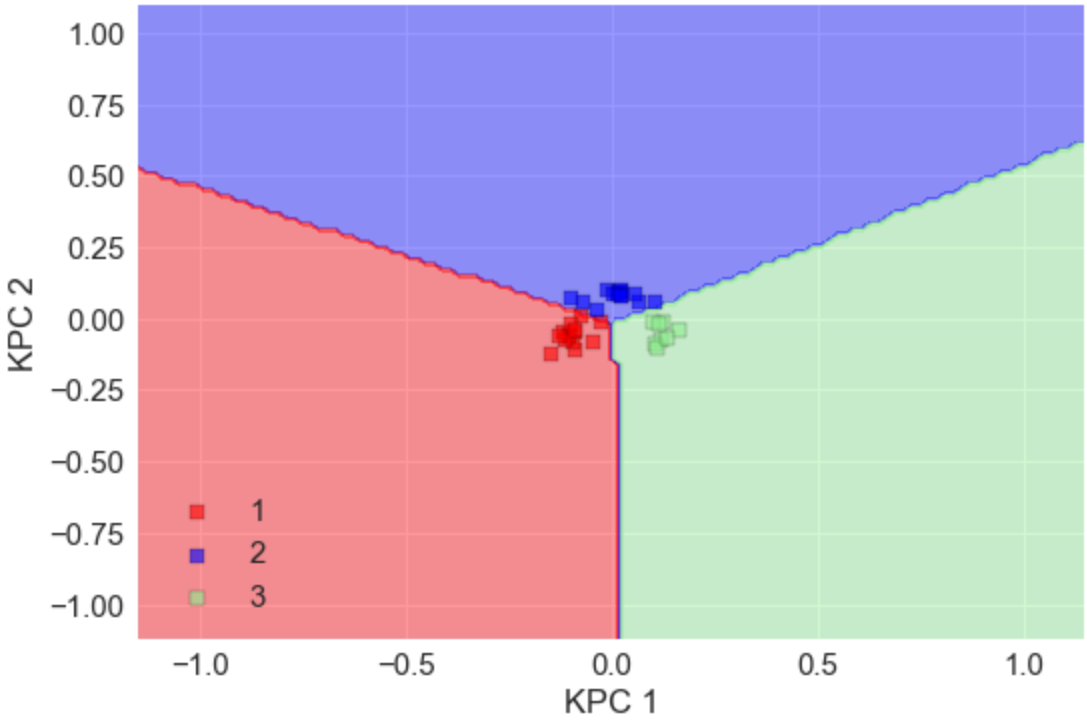


gamma 0.001



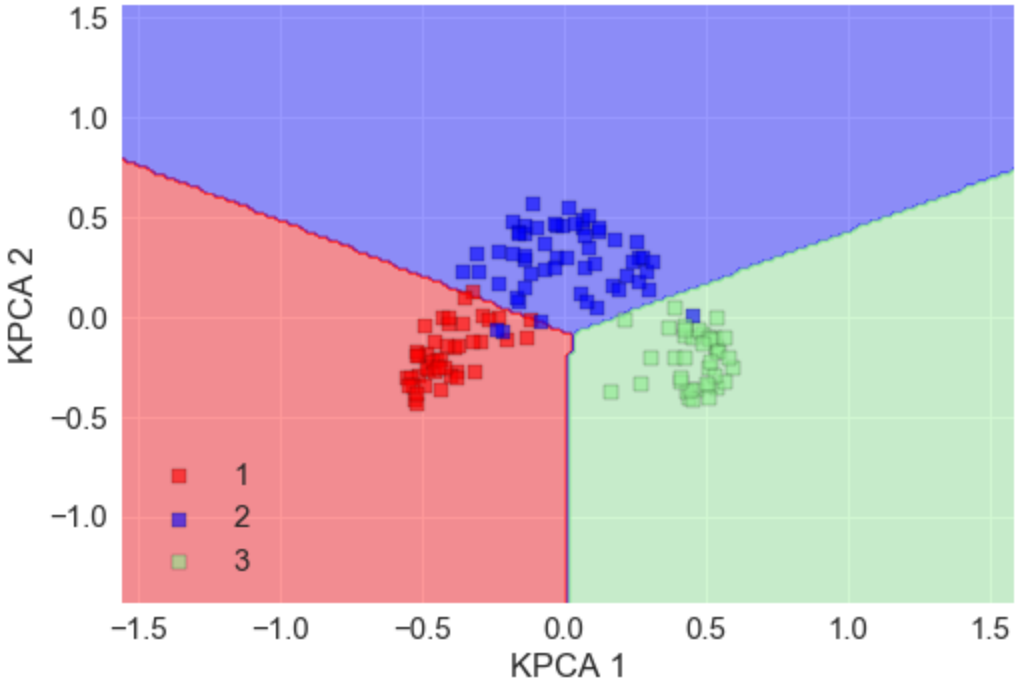
kpca sv train R^2: 0.96

```
gamma 0.001
```



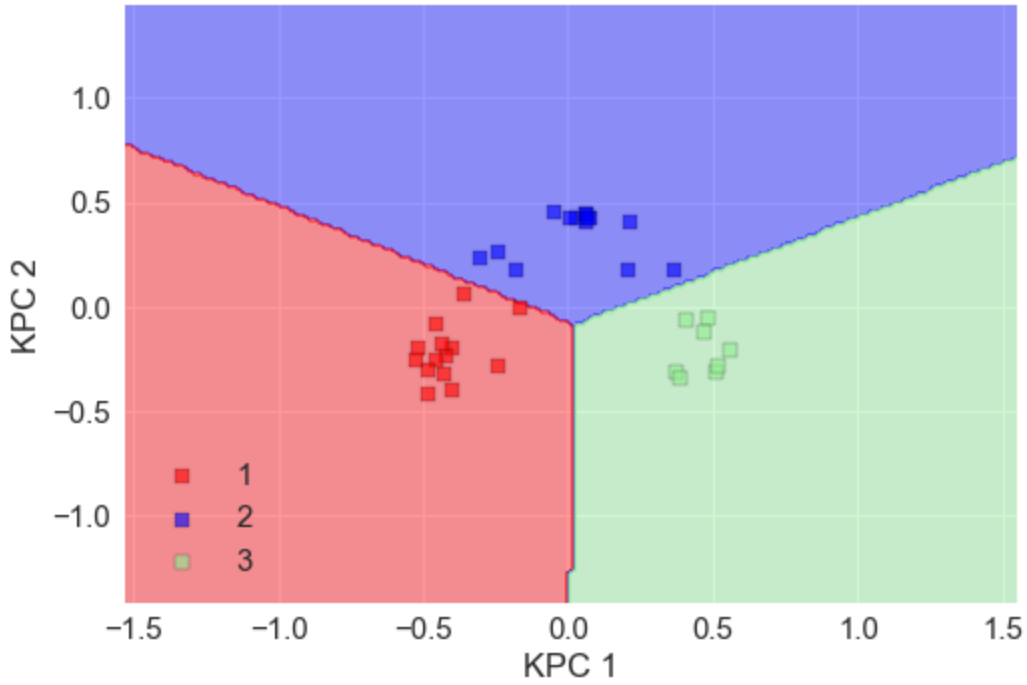

```
kpca sv test R^2: 1.00
```

```
gamma 0.0316227766017
```



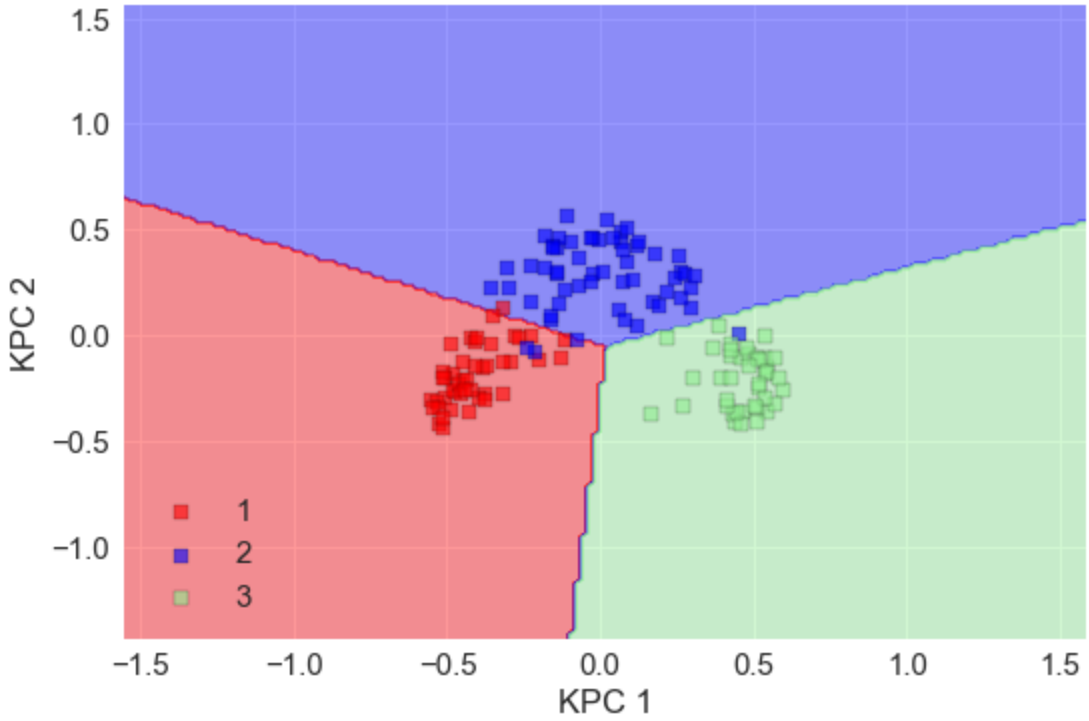

```
kpca lr trainR^2: 0.97
```

```
gamma 0.0316227766017
```



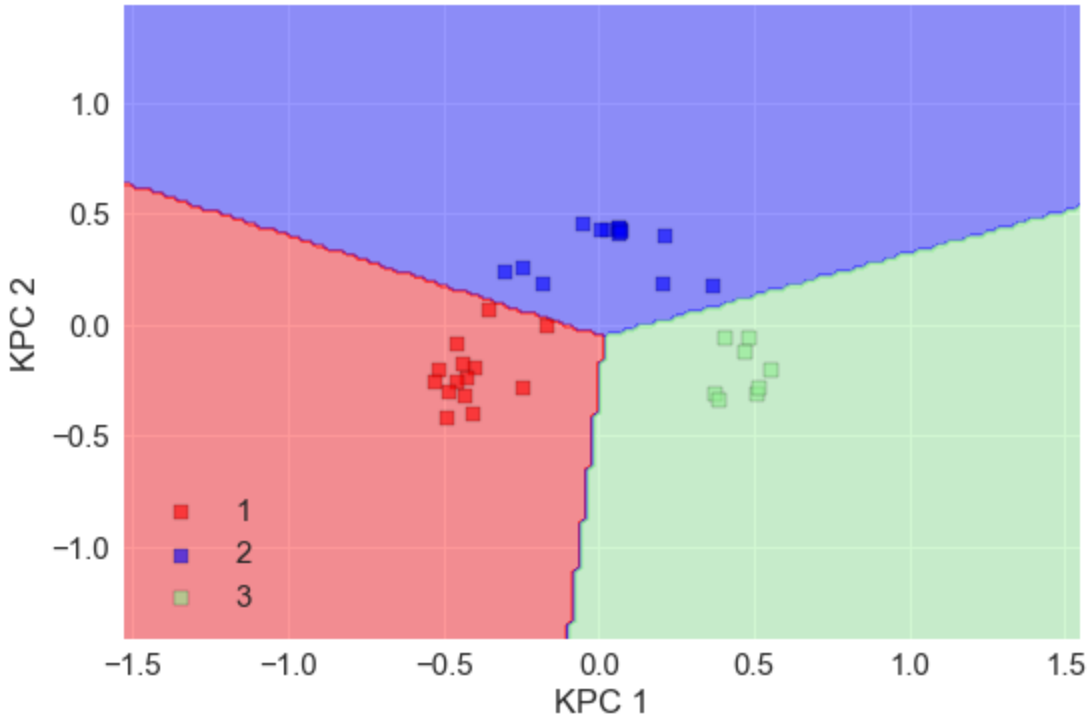

```
kpca lr test R^2: 1.00
```

```
gamma 0.0316227766017
```



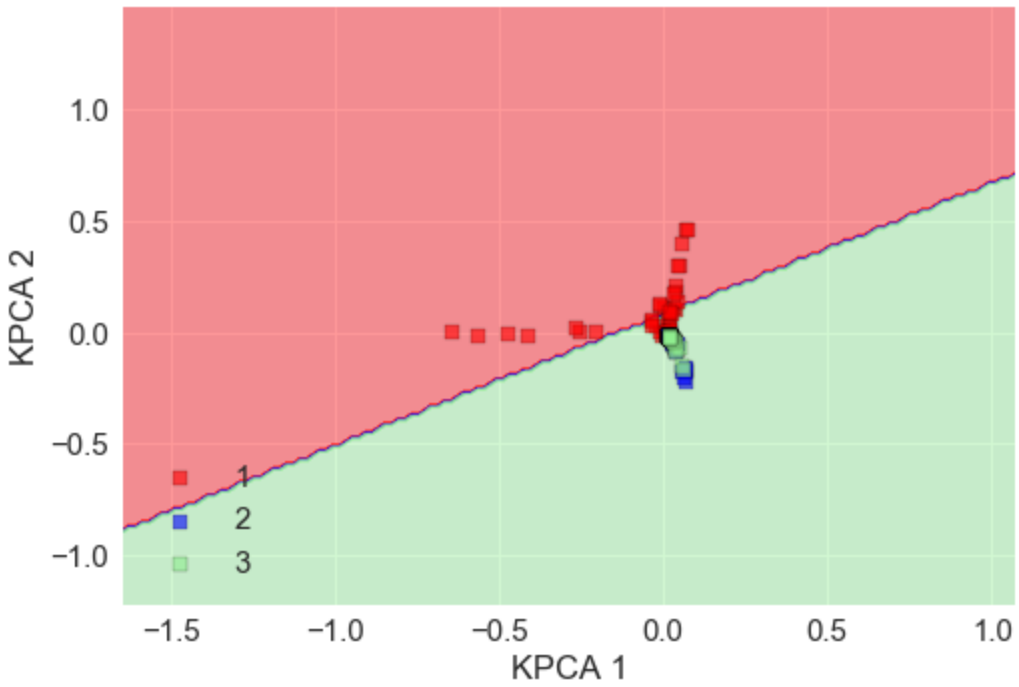

```
kpca sv train R^2: 0.96
```

gamma 0.0316227766017



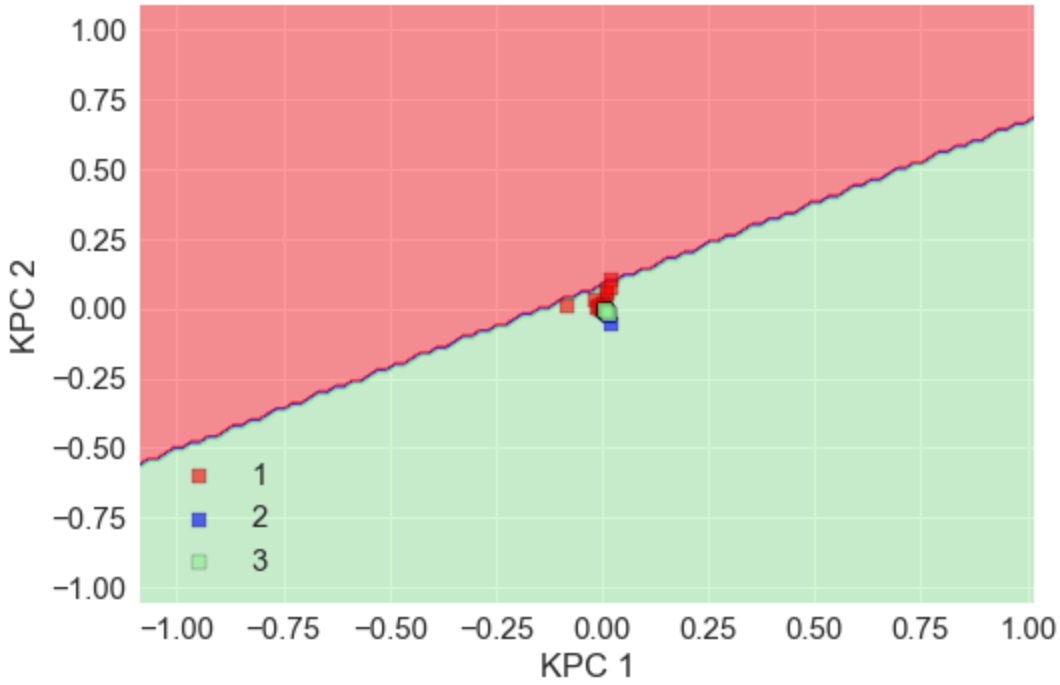
kpca sv test R^2: 1.00

gamma 1.0



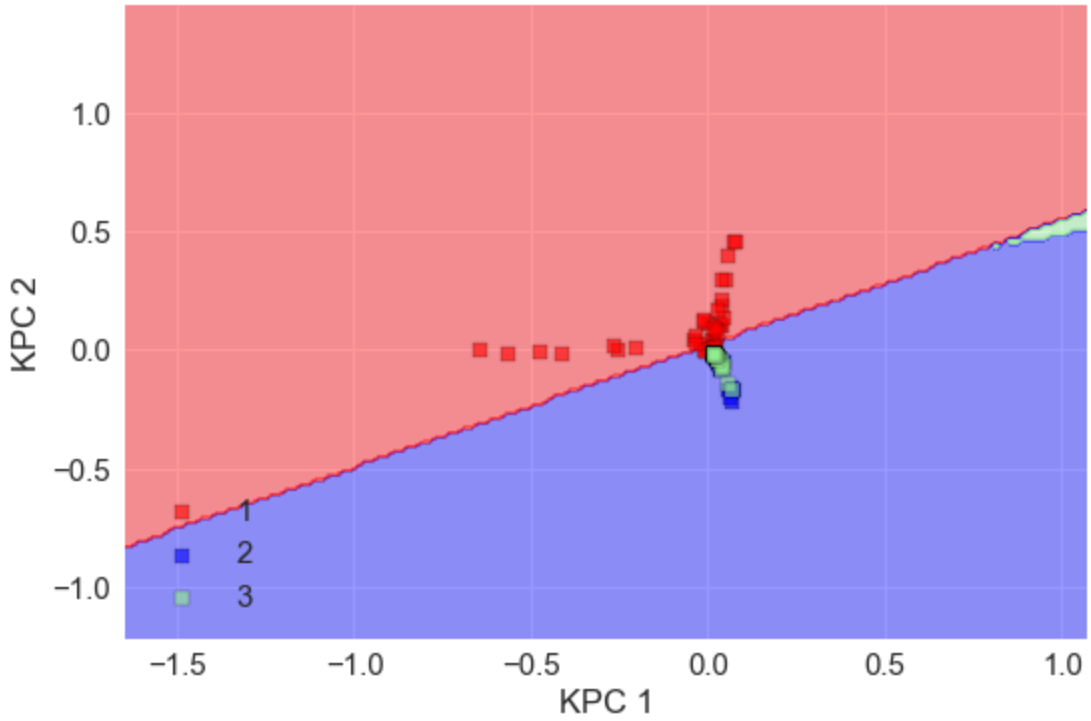
kpca lr trainR^2: 0.56

gamma 1.0



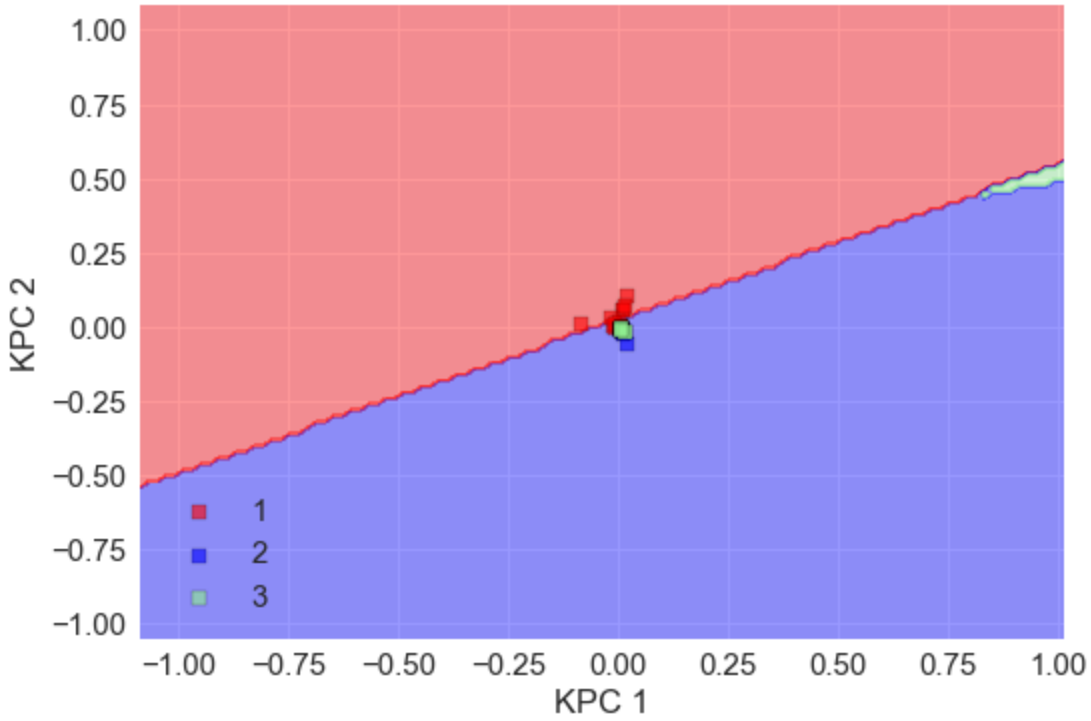
kpca lr test R^2: 0.42

gamma 1.0



kpca sv train R^2: 0.61

```
gamma 1.0
```




```
kpca sv test R^2: 0.56
```

Part 6: Conclusions

Write a short paragraph summarizing your findings. Which model performs best on the untransformed data? Which transformation leads to the best performance increases? Report your results using the Results worksheet format. Embed the completed table in your report.

Baseline	Train Acc: 1 Test Acc: 1	Train Acc: 0.99 Test Acc: 0.97
PCA transform	Train Acc: 0.96 Test Acc: 1	Train Acc: 0.96 Test Acc: 1
LDA transform	Train Acc: 0.99 Test Acc: 1	Train Acc: 0.99 Test Acc: 1

kPCA transform	Train Acc: 0.84 Test Acc: 0.89	Train Acc: 0.96 Test Acc: 1
Baseline	Train Acc: 1 Test Acc: 1	Train Acc: 0.99 Test Acc: 0.97
PCA transform	Train Acc: 0.96 Test Acc: 1	Train Acc: 0.96 Test Acc: 1
LDA transform	Train Acc: 0.99 Test Acc: 1	Train Acc: 0.99 Test Acc: 1
kPCA transform	Train Acc: 0.84 Test Acc: 0.89	Train Acc: 0.96 Test Acc: 1

From the picture, it is apparently that logistic model is performed well because the accuracy scores of train set and test set are both high. In fact, the data doesn't need to use other transformation because the accuracy is high enough. The PCA transformation is highest for the score.

Part 7: Appendix

Link to github repo

https://github.com/johnfeng123/Biao-Feng