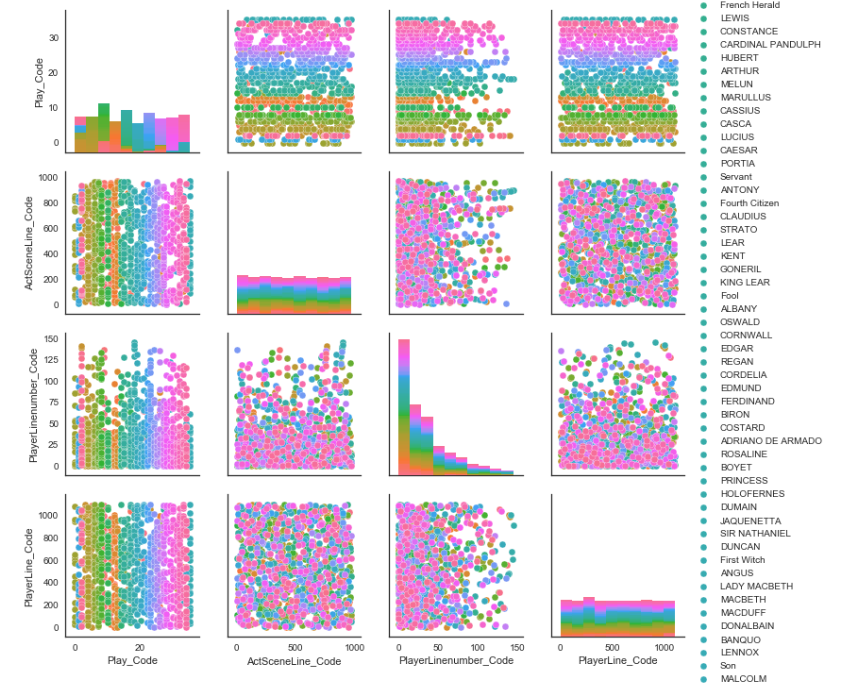
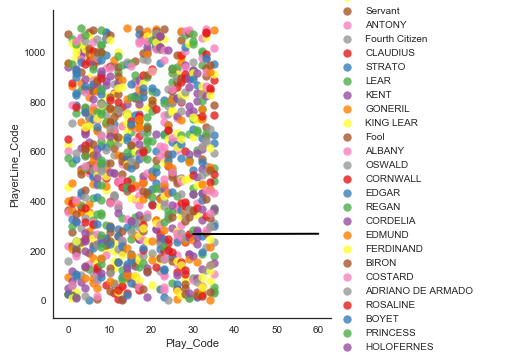
**Report**

The dataset used for this project held information regarding some of Shakespeare’s plays Henry IV and Henry VI part I. More specifically, the data set contained attributes such as play, the act scene line, the player line number, the play, and the player line. The purpose of this project was to determine which features in the dataset help to most accurately predict the correct player.

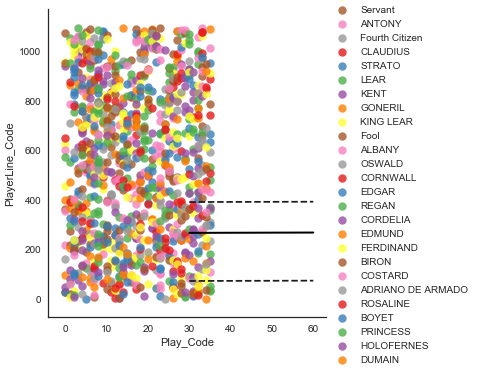
First, the data was inspected to determine which features would prove to be the most useful in helping to predict the correct player. To do so, a pair plot was constructed to see the correlation between the different features.



From the graph above, it is seen that the best separations appear among the play and the act scene line, the play and the player line number, and the play and the player line. The relationship between the play and player line is arbitrarily chosen because this appears to be the best and most equal separation of data. For this reason, these features will be the only ones used in the training and test sets needed for the construction of the linear support vector classification model. The support vector machine classifier was chosen arbitrarily because of its visualizations are fairly easy to understand. The visualization can be seen below.



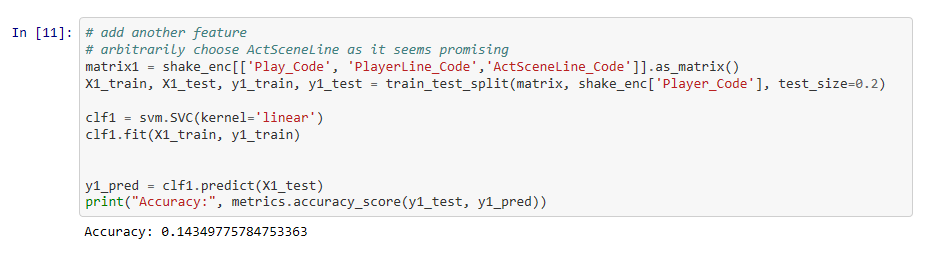
Below is the SVM visualization again with support vectors.



As seen by the graphs, the support vector machine did not provide a very good classification among the different players in the plays. As expected, the code supports this, as the model only had a prediction accuracy of around 13%.



Due to the poor accuracy of the first model, a second model was created adding another feature, the ActSceneLine, in hopes of substantially improving the prediction accuracy. However, the accuracy only improved somewhat.



Due to these disappointing results, perhaps this dataset should be analyzed using different classifier models rather than the support vector machine.