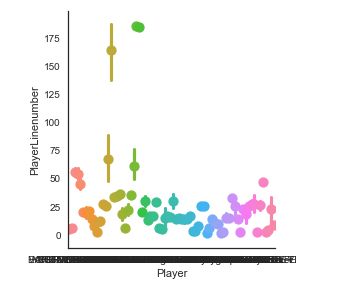
**Report**

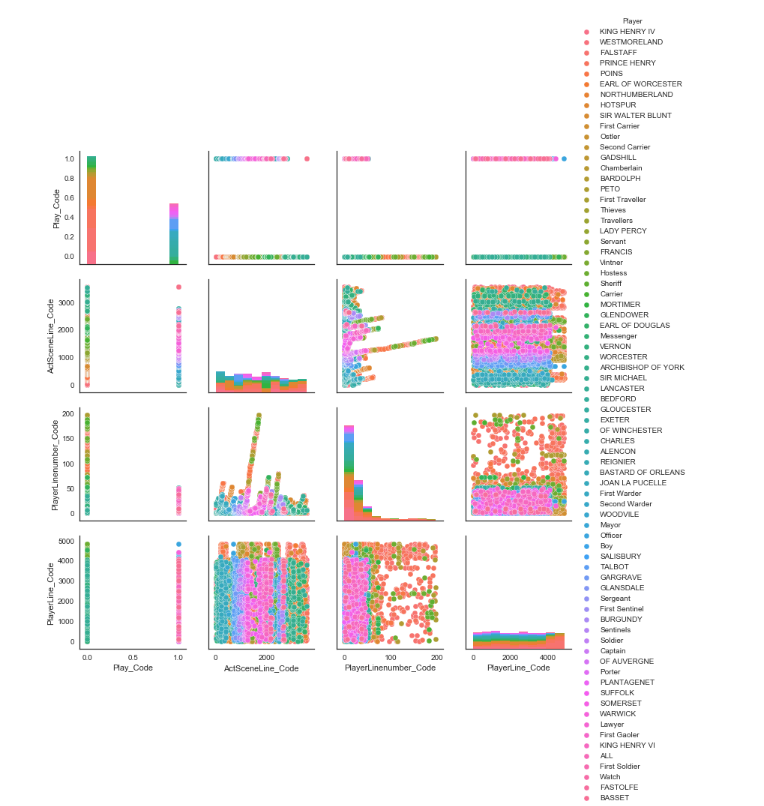
The dataset used for this project held information regarding some of Shakespeare’s plays Henry IV and Henry VI part I. More specifically, the data set contained attributes such as play, the act scene line, the player line number, the play, and the player line. The purpose of this project was to determine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

First, the data was plotted using categorical models, since the features is categorical.



However, this did not provide much useful information, so the categorical data was transformed into numerical data using label encoding. This allowed for easier modeling of the data and led to more useful information being extracted from the models.

Next, 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 can be 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 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. However, there was difficulty constructing this model.