Title: Predicting shot results from different areas in the NBA
Team members: Tomer Zur and Francis Rodriguez

Project Description: We are trying to predict whether a shot will go in using information such as the player who took the shot, where the shot was attempted on the court, the amount of time left, and the type of shot that was attempted. We would also like to incorporate players' historic shot percentages from each area on the court into these predictions.

Dataset: The dataset is linked here. The set has a record for each shot from the NBA (in the 2015-16 season), and it has many features such as where the player was on the court, how close the nearest defender was, the type of shot (i.e. a pull up jumper or a layup), and their opponent. There are between 84,000 and 85,000 records in the set. Features are primarily strings, but can be encoded into numbers for analysis with some basic data manipulation (i.e. coding a 1 for shot made and a 0 for shot missed, or a different number for each type of shot).

Approach and Methodology:

- Normalization: Lots of these variables are binary or categorical, and since the main numerical one is where the players are in, it's bounded by court size. Because of this, we will check the data for outliers before we normalize it, but it may not be necessary.
- Feature selection: We will likely drop some features that are not associated with the shot being made but are simply used for description (like the numerical game ID).
- Models: We will try to use a kNN model to classify the shots based on how likely shots around them were to have gone in, and we can also try to use other models like Logistic Regression or Naive Bayes to compare for the best model.
- Methodology for splitting: We will do a random test train split to split the existing data into approximately 75-80 percent training and 20-25 percent testing, and we can also cross validate the data using 5 to 10 k-folds.
- Language and packages: Python, we plan to use Scikit-Learn, numpy, matplotlib, plotly, and maybe tensorflow/keras.
- Metrics: We will evaluate our model using accuracy (% of shots correctly predicted / total shots) and f1 score (so our model doesn't overpredict makes or misses). We will also run cross validation to make sure our model isn't predicting way differently for our training data and testing data.