Plan Change:
The original plan for this project was to find a correlation between collegiate performances and their professional careers. This has since been narrowed down to collegiate basketball statistics correlating to rookie year NBA statistics and whether there is a large differential between the gap of collegiate play versus professional play. The years this dataset will span will be from the 2013-2020 NBA and collegiate seasons. We also want to be able to classify both the nba rookie and collegiate dataset using the kNN classifier.

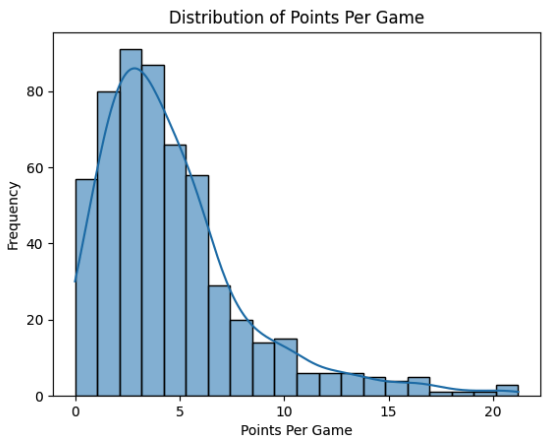
Data:
The training data for this project will be from the years 2013-2016, while the testing data will be from 2016-2020. This provides a seven year data set in which the changes of playstyle and decisions is minimal for both the NBA and collegiate levels of play. One of the problems with data in the NBA is the changes of playstyle from era to era, which leaves only one extended period of time where the change in playstyle was not as apparent. The 2004-2005 season marks the end of the deadball era, while the 2012-2013 season is one of the last seasons before the three point revolution. The 2007 season marks a transition for collegiate basketball where the elite collegiate basketball players decide to declare for the NBA draft rather than return for their sophomore season. This influx in the amount of one-and-done players would impact their NBA rookie seasons as well as their entire collegiate career due to the age and experience of these players. The seasons from 2013-2020 were chosen because the style of play remained relatively consistent, the amount of statistics available online, and the overlap of another respective collegiate basketball era.

Preprocessing on the Data:
A majority of the preprocessing performed on the data was to eliminate excessive players based on years outside our 2013-2019 range. The nba kaggle dataset was originally from the years 1997 - 2020 and the college dataset was from the range 2009 - 2021. We also wanted to determine whether there were any missing attributes, major outliers (such as players with nearly no games played and no statistically significant attributes), duplicates and other issues. An example is using dropna() which excludes missing values. None of these issues ever arose in our preprocessing of the kaggle datasets for either of the datasets. The way we viewed the data was using excel spreadsheets which also conveniently allowed filtering and other tools to clean the data. We were able to access the csv files through python and the pandas package which allowed us to begin our exploratory analysis of the datasets.

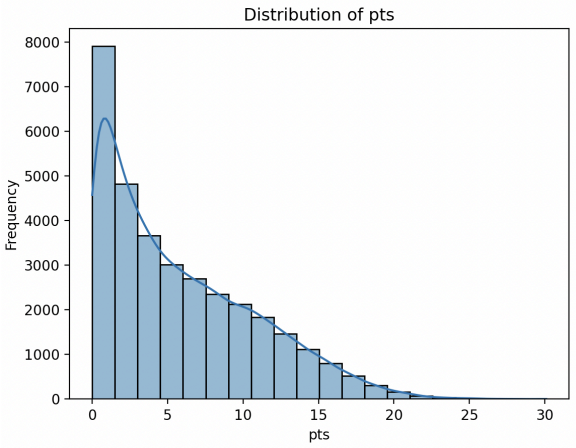
Exploratory Analysis:
Exploratory data analysis centered around two data sets: nba rookies statistics and college basketball players statistics. My first step in both analyses was to determine what kind of data was stored and in what form. Because our research is centered towards performance, I decided to emphasize these 14 statistics: Age, games, minutes played, points, field goal, field goal

percentage, 3 pointers, 3 point percentage, free throws, free throw percentage, total rebounds, assists, steals, and blocks. All of these variables are sorted as either floats or integers. For each of these variables, I was able to plot their distribution which I believe will help give us a better understanding of what performance can be considered good and also the trend of statistics among NBA and college players. Here is an example of points per game being plotted:

NBA:



Collegiate:



(This type of histogram analysis also exists for the other performance statistics.)

We initially have done a correlation analysis test between points per game for NBA rookies and college basketball players. We have gotten a -0.0350237 from the unprocessed datasets. Although there is still a lot of refinement necessary with both datasets, it appears that points per game is a poor indicator of collegiate to rookie performance in the NBA.

Performance:

So far, we have attempted to perform the kNN classification algorithm on the NBA rookies dataset with the test size equal to 0.3 of the dataset. We imported sklearn features to implement the processing and modeling in a python script. To standardize the features, we have used scaler.fit_transform on the train and the test. kNN was initialized to 3 neighbors. After running the python script we got an accuracy of 0.8683 which appears to be moderately successful. Other algorithms we may try are naive bayes or the decision tree (resist irrelevant attributes).

Next Steps:

Following steps will be to perform correlation analysis on other performance bearing statistics like rebounds, field goal percentage per game and others. All of these will need to be plotted and tracked so that we can determine the best statistic for success in college as well as the NBA. Secondly, we need to refine the kNN script for the NBA dataset and perform the kNN algorithm on the college dataset. This involves analyzing the underfitting or overfitting as well as continuing to refine the dataset for irrelevant attributes. So far we have excluded unnamed players, the player name attribute, team, year, conference, and target in our preprocessing.