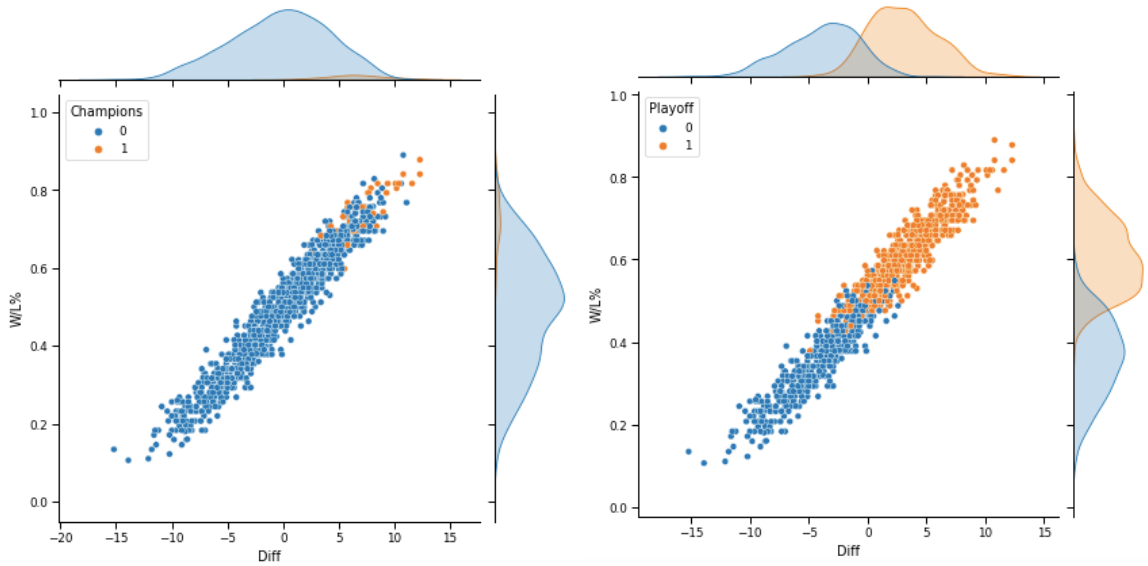
I really enjoy the NBA and wanted to explore how teams over or underperform based on their record at the end of the season and then use this sort of data to predict playoff/champion teams in the 2020-2021 season. To do this, I web scraped data from basketball-reference.com using Python from 1972 to 2019, which resulted in a dataframe with the following data:

| | Team | W | L | W/L% | GB | PS/G | PA/G | SRS | Year | Playoff | Diff | Champions |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Cleveland Cavaliers* | 57 | 25 | 0.695 | — | 104.3 | 98.3 | 5.45 | 2016 | 1 | 6.0 | 1 |
| **1** | Toronto Raptors* | 56 | 26 | 0.683 | 1.0 | 102.7 | 98.2 | 4.08 | 2016 | 1 | 4.5 | 0 |
| **2** | Miami Heat* | 48 | 34 | 0.585 | 9.0 | 100.0 | 98.4 | 1.50 | 2016 | 1 | 1.6 | 0 |
| **3** | Atlanta Hawks* | 48 | 34 | 0.585 | 9.0 | 102.8 | 99.2 | 3.49 | 2016 | 1 | 3.6 | 0 |
| **4** | Boston Celtics* | 48 | 34 | 0.585 | 9.0 | 105.7 | 102.5 | 2.84 | 2016 | 1 | 3.2 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **11** | Portland Trail Blazers | 41 | 41 | 0.500 | 15.0 | 90.7 | 92.0 | -0.58 | 2004 | 0 | -1.3 | 0 |
| **12** | Golden State Warriors | 37 | 45 | 0.451 | 19.0 | 93.3 | 94.0 | -0.07 | 2004 | 0 | -0.7 | 0 |
| **13** | Seattle SuperSonics | 37 | 45 | 0.451 | 19.0 | 97.1 | 97.8 | 0.02 | 2004 | 0 | -0.7 | 0 |
| **14** | Phoenix Suns | 29 | 53 | 0.354 | 27.0 | 94.2 | 97.9 | -2.94 | 2004 | 0 | -3.7 | 0 |
| **15** | Los Angeles Clippers | 28 | 54 | 0.341 | 28.0 | 94.8 | 99.4 | -3.74 | 2004 | 0 | -4.6 | 0 |

1257 rows × 12 columns

A lot of useful data here, including points scores per game (ps/g), points allowed per game (pa/g), the difference between those two (Diff), and whether that team won the championship (Champions) or went to the playoffs (Playoff). I wanted to focus on two variables: Diff and the win/less percentage (W/L%). Why? Those seem like good predictors of performance. If you score more than your opponent often, your diff will be high and your subsequence W/L would be high. Let's look at some jointplots of these variables along with grouping by Champions/Playoffs:




About what you would expect: champions and those that made the playoff are further to the top right. But it looks like my predictors could be simplified quite a bit. I used principle component analysis (PCA) to reduce my two dependent variables to just be one:

```
[9]  variables = ['W/L%','Diff']
     X = total_data[variables]
     y = total_data['Playoff']
```

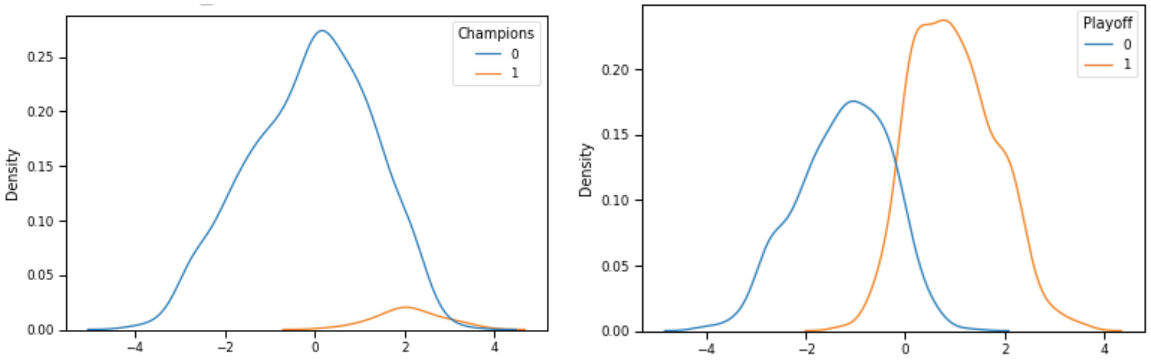
```
[10] scaler = StandardScaler()
     X_array = scaler.fit_transform(X)
     X2 = pd.DataFrame(X_array,columns=X.columns)

     pca = PCA()
     x_pca = pca.fit_transform(X2)
```

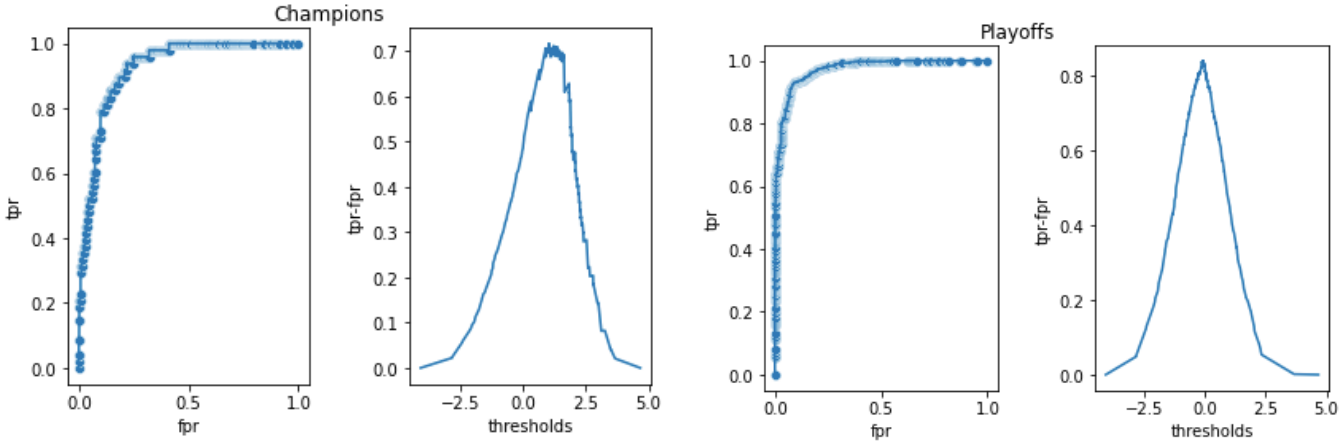
```
[ ]  pca.explained_variance_ratio_
```

```
     array([0.98410355, 0.01589645])
```

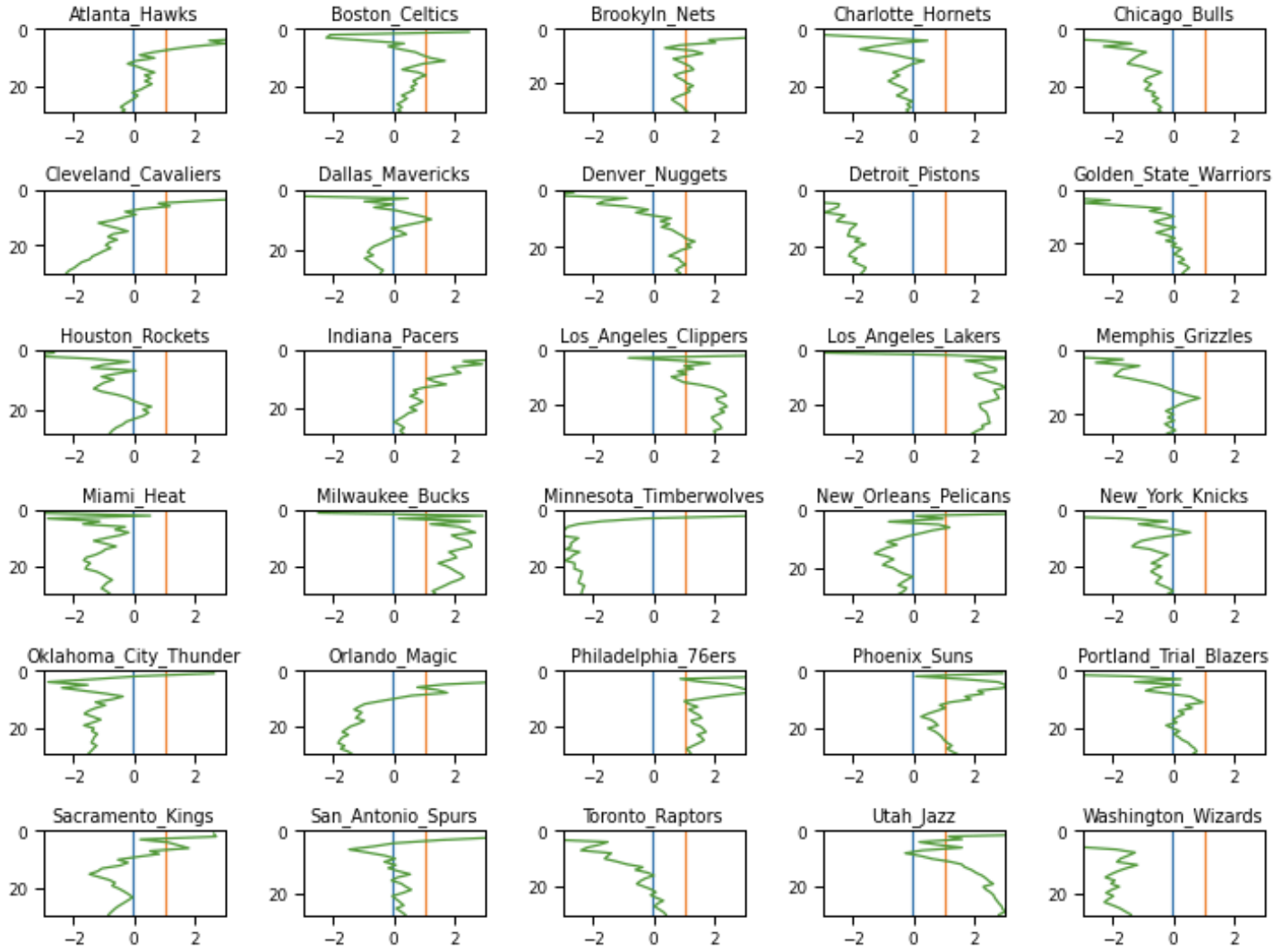
The first component explains much of the variance in the two variables. We can then look at our single component as a predictor of champion/playoff




Looks pretty good. I can then use this data to create a criterion for determining how likely a team will be a champion or playoff. I did this by looking at ROC and finding the maximum difference between the true positive rate (tpr) and false positive rate (fpr):




This creates criterions of 1.04 for champions and -0.06 for playoffs. Let's now get the boxscores for teams in the 2020-2021 season, calculate Diff and W/L% game-by-game, then pass those values into the PCA to see which teams are crossing these criterions:




Now, this is only scores of games up to 02/20/2021, so there is still quite a bit of variance. Here is a table with the current sorted end points (green highlighted teams are the teams to keep an eye on):

| | Team | Value |
|---|---|---|
| 28 | Utah_Jazz | 2.823211 |
| 12 | Los_Angeles_Clippers | 2.026494 |
| 13 | Los_Angeles_Lakers | 1.941472 |
| 23 | Phoenix_Suns | 1.460177 |
| 16 | Milwaukee_Bucks | 1.380021 |
| 22 | Philadelphia_76ers | 1.246214 |
| 2 | Brookyln_Nets | 1.140718 |
| 7 | Denver_Nuggets | 0.876857 |
| 24 | Portland_Trial_Blazers | 0.631318 |
| 27 | Toronto_Raptors | 0.408058 |
| 26 | San_Antonio_Spurs | 0.379293 |
| 11 | Indiana_Pacers | 0.315309 |
| 9 | Golden_State_Warriors | 0.309956 |
| 1 | Boston_Celtics | 0.272539 |
| 3 | Charlotte_Hornets | -0.133969 |
| 19 | New_York_Knicks | -0.176116 |
| 14 | Memphis_Grizzles | -0.202961 |

| | | |
|---|---|---|
| 4 | Chicago_Bulls | -0.364050 |
| 0 | Atlanta_Hawks | -0.428399 |
| 6 | Dallas_Mavericks | -0.437853 |
| 18 | New_Orleans_Pelicans | -0.476515 |
| 15 | Miami_Heat | -0.791298 |
| 10 | Houston_Rockets | -0.798170 |
| 25 | Sacramento_Kings | -0.834325 |
| 21 | Orlando_Magic | -1.365136 |
| 29 | Washington_Wizards | -1.396311 |
| 20 | Oklahoma_City_Thunder | -1.519224 |
| 8 | Detroit_Pistons | -1.717616 |
| 5 | Cleveland_Cavaliers | -2.228571 |
| 17 | Minnesota_Timberwolves | -2.384173 |

The next thing to do with this data, other than track 2020-2021 season, is to see which teams most over and under performed between 1972-2019. Based on the residuals of the Diff and W/L% from the scatterplots above, see below for the under and over achievers:

```
[ ] total_data[total_data['Residuals2'] < -.10268227] # underachievers
```

| | Team | W | L | W/L% | GB | PS/G | PA/G | SRS | Year | Diff | Champions | Residuals2 | P-W/L% | P-W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **12** | Dallas Mavericks | 24 | 58 | 0.293 | 41.0 | 102.3 | 105.4 | -2.70 | 2018 | -3.1 | 0 | -0.107036 | 0.400036 | 32.802922 |
| **3** | Philadelphia 76ers* | 35 | 31 | 0.530 | 4.0 | 93.6 | 89.4 | 3.59 | 2012 | 4.2 | 0 | -0.105348 | 0.635348 | 41.932990 |
| **8** | Houston Rockets | 32 | 50 | 0.390 | 15.0 | 107.4 | 107.6 | -0.34 | 1974 | -0.2 | 0 | -0.103516 | 0.493516 | 40.468314 |
| **4** | Chicago Bulls | 24 | 58 | 0.293 | 14.0 | 95.9 | 98.8 | -2.89 | 1976 | -2.9 | 0 | -0.113483 | 0.406483 | 33.331570 |
| **12** | Phoenix Suns | 34 | 48 | 0.415 | 19.0 | 104.9 | 104.2 | 0.64 | 1977 | 0.7 | 0 | -0.107527 | 0.522527 | 42.847229 |
| **3** | Milwaukee Bucks | 38 | 44 | 0.463 | 10.0 | 114.1 | 111.8 | 2.12 | 1979 | 2.3 | 0 | -0.111103 | 0.574103 | 47.076410 |
| **12** | Seattle SuperSonics | 31 | 51 | 0.378 | 31.0 | 104.4 | 104.5 | -0.47 | 1986 | -0.1 | 0 | -0.118739 | 0.496739 | 40.732638 |

```
[ ] total_data[total_data['Residuals2'] > 0.09600996] # overachievers
```

| | Team | W | L | W/L% | GB | PS/G | PA/G | SRS | Year | Diff | Champions | Residuals2 | P-W/L% | P-W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **7** | Golden State Warriors* | 51 | 31 | 0.622 | 18.0 | 108.2 | 107.4 | 0.92 | 1972 | 0.8 | 0 | 0.096249 | 0.525751 | 43.111553 |
| **2** | Boston Celtics* | 44 | 38 | 0.537 | 6.0 | 104.5 | 106.5 | -1.90 | 1977 | -2.0 | 0 | 0.101506 | 0.435494 | 35.710484 |
| **11** | Los Angeles Clippers | 32 | 50 | 0.390 | 30.0 | 108.6 | 115.5 | -6.83 | 1986 | -6.9 | 0 | 0.112456 | 0.277544 | 22.758615 |
| **4** | Miami Heat* | 38 | 44 | 0.463 | 13.0 | 105.0 | 109.2 | -3.94 | 1992 | -4.2 | 0 | 0.098422 | 0.364578 | 29.895360 |
| **6** | Dallas Mavericks | 11 | 71 | 0.134 | 44.0 | 99.3 | 114.5 | -14.68 | 1993 | -15.2 | 0 | 0.124003 | 0.009997 | 0.819735 |
| **8** | Phoenix Suns* | 59 | 23 | 0.720 | — | 110.6 | 106.8 | 3.86 | 1995 | 3.8 | 0 | 0.097546 | 0.622454 | 51.041268 |
| **12** | Charlotte Hornets* | 54 | 28 | 0.659 | 15.0 | 98.9 | 97.0 | 2.13 | 1997 | 1.9 | 0 | 0.097791 | 0.561209 | 46.019115 |

The next step is to go game-by-game and see what happened. Was it luck? Stay tuned!