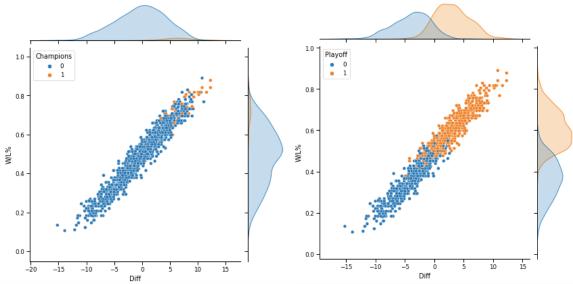
I really enjoy the NBA and wanted to explore how teams over or underperform based on their record at the end of the season and then use this sort of data to predict playoff/champion teams in the 2020-2021 season. To do this, I web scraped data from basketball-reference.com using Python from 1972 to 2019,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

Team	W	L	W/L%	GB	PS/G	PA/G	SRS	Year	Playoff	Diff	Champions
I Cavaliers*	57	25	0.695	-	104.3	98.3	5.45	2016	1	6.0	1
to Raptors*	56	26	0.683	1.0	102.7	98.2	4.08	2016	1	4.5	0
/liami Heat*	48	34	0.585	9.0	100.0	98.4	1.50	2016	1	1.6	0
nta Hawks*	48	34	0.585	9.0	102.8	99.2	3.49	2016	1	3.6	0
ton Celtics*	48	34	0.585	9.0	105.7	102.5	2.84	2016	1	3.2	0
rail Blazers	41	41	0.500	15.0	90.7	92.0	-0.58	2004	0	-1.3	0
te Warriors	37	45	0.451	19.0	93.3	94.0	-0.07	2004	0	-0.7	0
uperSonics	37	45	0.451	19.0	97.1	97.8	0.02	2004	0	-0.7	0
oenix Suns	29	53	0.354	27.0	94.2	97.9	-2.94	2004	0	-3.7	0
es Clippers	28	54	0.341	28.0	94.8	99.4	-3.74	2004	0	-4.6	0
֡֡֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜֜	I Cavaliers* to Raptors* Miami Heat* nta Hawks* ton Celtics* Trail Blazers ate Warriors uperSonics	to Raptors* 57  to Raptors* 56  Miami Heat* 48  Inta Hawks* 48  ton Celtics* 48  Inta Blazers 41  Inte Warriors 37  uperSonics 37  openix Suns 29	## Cavaliers* 57 25  ## to Raptors* 56 26  ## ## ## ## ## ## 34  ## mta Hawks* 48 34  ## ton Celtics* 48 34  ##  ## trail Blazers 41 41  ## trail Warriors 37 45  ## uperSonics 37 45  ## toological colors and selections are selected as a selection and selection are selected as a selection and selection are selected as a selection and selection are selected as a selection are selec	## Cavaliers* 57 25 0.695  ## to Raptors* 56 26 0.683  ## Miami Heat* 48 34 0.585  ## nta Hawks* 48 34 0.585  ## ton Celtics* 48 34 0.585  ## ton Celtics* 48 34 0.585  ## in a Hawks* 48 34 0.585  ##	to Raptors* 57 25 0.695 —  to Raptors* 56 26 0.683 1.0  Miami Heat* 48 34 0.585 9.0  Inta Hawks* 48 34 0.585 9.0  ton Celtics* 48 34 0.585 9.0   Trail Blazers 41 41 0.500 15.0  Inte Warriors 37 45 0.451 19.0  Inte Source Source 37 45 0.451 19.0  Interpresentation of the property of	I Cavaliers*       57       25       0.695       —       104.3         Ito Raptors*       56       26       0.683       1.0       102.7         Miami Heat*       48       34       0.585       9.0       100.0         Inta Hawks*       48       34       0.585       9.0       102.8         Iton Celtics*       48       34       0.585       9.0       105.7         Irail Blazers       41       41       0.500       15.0       90.7         Ite Warriors       37       45       0.451       19.0       93.3         IuperSonics       37       45       0.451       19.0       97.1         Ioenix Suns       29       53       0.354       27.0       94.2	I Cavaliers*       57       25       0.695       —       104.3       98.3         Ito Raptors*       56       26       0.683       1.0       102.7       98.2         Miami Heat*       48       34       0.585       9.0       100.0       98.4         Inta Hawks*       48       34       0.585       9.0       102.8       99.2         Iton Celtics*       48       34       0.585       9.0       105.7       102.5         Interview       41       41       0.500       15.0       90.7       92.0         Ite Warriors       37       45       0.451       19.0       93.3       94.0         Interview       37       45       0.451       19.0       97.1       97.8         Interview       38       34       0.354       27.0       94.2       97.9	Cavaliers* 57 25 0.695 — 104.3 98.3 5.45 to Raptors* 56 26 0.683 1.0 102.7 98.2 4.08 Miami Heat* 48 34 0.585 9.0 100.0 98.4 1.50 Inta Hawks* 48 34 0.585 9.0 102.8 99.2 3.49 ton Celtics* 48 34 0.585 9.0 105.7 102.5 2.84 Interval Blazers 41 41 0.500 15.0 90.7 92.0 -0.58 Interval Warriors 37 45 0.451 19.0 93.3 94.0 -0.07 uperSonics 37 45 0.451 19.0 97.1 97.8 0.02 openix Suns 29 53 0.354 27.0 94.2 97.9 -2.94	Cavaliers* 57 25 0.695 — 104.3 98.3 5.45 2016 to Raptors* 56 26 0.683 1.0 102.7 98.2 4.08 2016 Miami Heat* 48 34 0.585 9.0 100.0 98.4 1.50 2016 Inta Hawks* 48 34 0.585 9.0 102.8 99.2 3.49 2016 ton Celtics* 48 34 0.585 9.0 105.7 102.5 2.84 2016 Inta Blazers 41 41 0.500 15.0 90.7 92.0 -0.58 2004 Inte Warriors 37 45 0.451 19.0 93.3 94.0 -0.07 2004 Interpretation Celtics* 29 53 0.354 27.0 94.2 97.9 -2.94 2004	1 Cavaliers* 57 25 0.695 — 104.3 98.3 5.45 2016 1 1 to Raptors* 56 26 0.683 1.0 102.7 98.2 4.08 2016 1 1 Miami Heat* 48 34 0.585 9.0 100.0 98.4 1.50 2016 1 1 ton Celtics* 48 34 0.585 9.0 102.8 99.2 3.49 2016 1 1 ton Celtics* 48 34 0.585 9.0 105.7 102.5 2.84 2016 1 1 mia Blazers 41 41 0.500 15.0 90.7 92.0 -0.58 2004 0 1 tel Warriors 37 45 0.451 19.0 93.3 94.0 -0.07 2004 0 1 uperSonics 37 45 0.451 19.0 97.1 97.8 0.02 2004 0 1 openix Suns 29 53 0.354 27.0 94.2 97.9 -2.94 2004 0	Cavaliers* 57 25 0.695 — 104.3 98.3 5.45 2016 1 6.0 to Raptors* 56 26 0.683 1.0 102.7 98.2 4.08 2016 1 4.5 Miami Heat* 48 34 0.585 9.0 100.0 98.4 1.50 2016 1 1.6 Inta Hawks* 48 34 0.585 9.0 102.8 99.2 3.49 2016 1 3.6 ton Celtics* 48 34 0.585 9.0 105.7 102.5 2.84 2016 1 3.2 Inta Blazers 41 41 0.500 15.0 90.7 92.0 -0.58 2004 0 -1.3 ate Warriors 37 45 0.451 19.0 93.3 94.0 -0.07 2004 0 -0.7 uperSonics 37 45 0.451 19.0 97.1 97.8 0.02 2004 0 -0.7 doenix Suns 29 53 0.354 27.0 94.2 97.9 -2.94 2004 0 -3.7

1257 rows x 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:

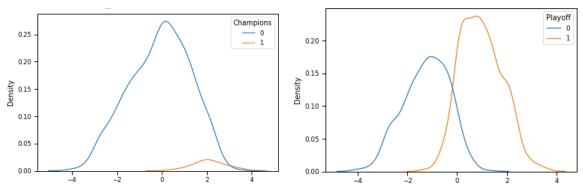
```
[9] variables = ['W/L%','Diff']
   X = total_data[variables]
   y = total_data['Playoff']

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

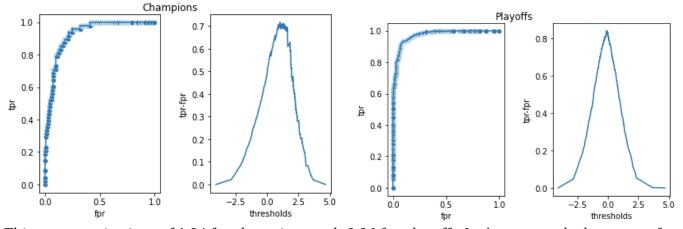
   pca = PCA()
   x_pca = pca.fit_transform(X2)

[ ] pca.explained_variance_ratio_
   array([0.98410355, 0.01589645])
```

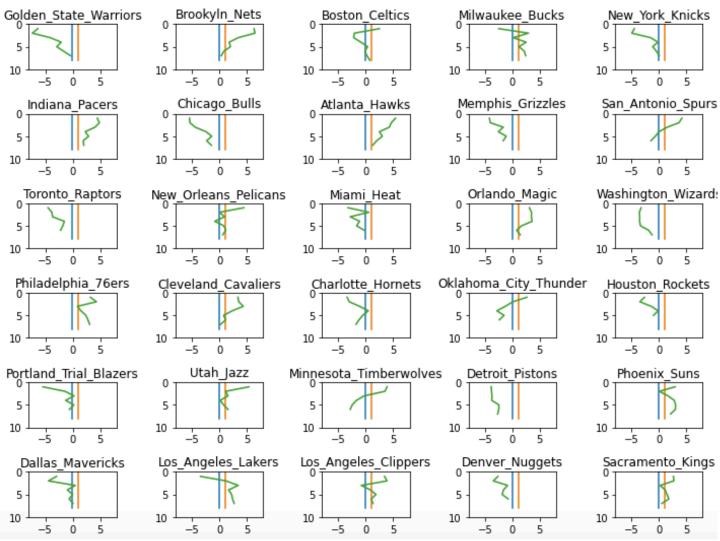
The first component explains much of the variance in the two variables. We can then look at our single component as a predictor of champion/playoff



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 01/04/2021, so there is still quite a bit of variance. However, up to this point, its pretty clear the Lakers, 76ers, Bucks, and Suns all seem to be trending in the right direction, which more or less matches the eye-test.

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:

tota	al_data[total_data	a['F	Resi	duals2	] < .	10268	227] #	under	achiev	ers					
	Team	W	L	W/L%	GB	PS/G	PA/G	SRS 3	lear D	iff C	hampions	Resid	luals2	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	21	E4	0.070		4044	4045	0.47	4000	-0.1	0		440700	0.406700	40.732638
12	Seattle Supersonics	31	51	0.378	31.0	104.4	104.5	-0.47	1986	-0.1	0	-0.	118/39	0.496739	40.732638
	al_data[total_data	a['F	Resi	duals2	] > (	0.09600	996]#	overa	chieve	rs					
tota	al_data[total_data	a['F	Resi W	duals2 L W/I	'] > (	0.09600 B PS/	996] # G PA/G	overa	chieve	rs Diff	Champio	ns Re	sidual	s2 P-W/	L% F
	al_data[total_data	a['F m s* 5	W 51 (	duals2	'] > ( % G	0.09600 B PS/	996] # <b>G PA/G</b> 2 107.4	overa	chievers S Year 2 1972	Diff	Champio			s <b>2 P-</b> W/	L% F
tota	al_data[total_data Tea Golden State Warriors	m s* 5	W 51 3	duals2 L W/I 31 0.62	'] > ( % G 2 18.	0.09600 B PS/0 0 108.3	996] # <b>G PA/G</b> 2 107.4 5 106.5	overa	chiever <b>S Year</b> 2 1972 0 1977	Diff 0.8	Champio	ns Re	e <b>sidual</b> 0.0962	s2 P-W/ 49 0.5257 06 0.4354	L% F 751 43.1115
7 2	al_data[total_data Tea  Golden State Warriors  Boston Celtics	m. s* 5 s* 4	W 51 3	duals2 L W/L 31 0.62	G2 18.67 6.60 30.	0.09600 B PS/6 0 108.6 0 104.6	996] # <b>G PA/G</b> 2 107.4 5 106.5 6 115.5	overa  SR3  0.93  -1.96  -6.88	chiever <b>S Year</b> 2 1972 0 1977 3 1986	Diff 0.8 -2.0 -6.9	Champio	ns Re	0.0962	s2 P-W/49 0.5257 06 0.4354 56 0.2775	L% F 751 43.1115 194 35.7104 544 22.7586
7 2	al_data[total_data Tea  Golden State Warriors Boston Celtics Los Angeles Clipper	m. s* 5 s* 4 s* 3	W 51 31 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	duals2 L W/I 31 0.62 38 0.53 50 0.39	"] > 0 % G 22 18. 7 6. 90 30.	0.09600 B PS/0 0 108.3 0 104.3 0 108.6 0 105.6	996] # <b>G PA/G</b> 2 107.4  5 106.5  6 115.5  0 109.2	overa	chiever s Year 2 1972 0 1977 3 1986 4 1992	Difff 0.8 -2.0 -6.9	Champio	ns Re 0 0	0.0962 0.1015 0.1124	s2 P-W/ 49 0.5257 06 0.4354 56 0.2775 22 0.3645	L% I 751 43.1115 194 35.710 194 22.758 198 29.895
7 2 11 4	al_data[total_data Tea  Golden State Warriors Boston Celtics Los Angeles Clipper Miami Hea	m	W	duals2 L W/I 31 0.62 38 0.53 50 0.39	G2 18.67 6.60 30.63 13.64 44.65	0.09600 B PS/0 0 108.3 0 104.3 0 108.6 0 105.6	996] #  G PA/G  2 107.4  5 106.5  6 115.5  0 109.2  3 114.5	overa	chiever <b>S Year</b> 2 1972 0 1977 3 1986 4 1992 8 1993	Difff 0.8 -2.0 -6.9 -4.2	Champio	ns Re 0 0 0 0	0.0962 0.1015 0.1124 0.0984	s2 P-W/ 49 0.5257 06 0.4354 56 0.2775 22 0.3645 03 0.0098	L% F 751 43.1115 94 35.7104 644 22.7586 678 29.8953

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