Term Project 227

Team Member Name(s): Benjamin Mertens

Project Title:

What makes the best sum of individuals for the greatest whole in NBA team performance.

Problem statement:

The objective of this research was to determine the model of champions in the NBA by examining a variety of individual-level factors that contribute to the performance of the whole team. These factors included salary, player roster emphasis, player age, player continuity, and coaching effectiveness.

The first step in doing this was trying to rate players' individual impacts and cluster them into a player category. This was the most important step in the sense where it will colored the rest of the analysis. One way that could have been done to do this would have been to use advanced statistics of box plus-minus (BPM), win shares (WS), or player efficiency rating (PER). These are supposed to be all-in-one statistics for an individual's performance rated against their peers. Another way would be assigning weights to how far a team went in a playoff run and giving players value based off minutes played. The reason this would make sense is that it would show an individual's impact on winning throughout years, and championships are often the most important thing in the NBA. Michael Jordan is still widely considered the greatest of all time (GOAT) based off 6 championships in 8 years with his longevity. Having never lost a finals series is one of his best points versus his biggest rival in LeBron James in the GOAT debate. LeBron is considered in the talks of GOAT based off making the championship series 8 years in a row, making 10 total, but *ONLY* winning 4. So, with the hope that this research would determine how best to build a championship winning team there needed to be a combination of individual merit and team success.

To do this, players were classified into clusters based off of all the information that could be gathered on them from the season before, and how successful their teams were the five years previous. This included all the statistics that could be scraped as well as how far the team went in the playoffs, or even made them. The players would not be held down by notions of assigned positions that modern research has shown to be mostly archaic. This categorizing would not include teams, teammates, coach, how the team acquired the player, or when the team acquired the player. Another aspect that was looked at was the age make-up of a winning team, but this one had the predicted issue in that age is a critical aspect in categorizing the players initially. This happened with salary and player pick in the draft also. Once the players were classified the next steps in the research was to examine these factors.

Once the players were classified, the next steps were to look at other factors that might influence to a team success rather than individual performance alone. To build a proper team, a team needs to pay players. However, payment in the NBA is a team-building resource as there is a salary cap that is not present in teams like European Football leagues or Major Legue Baseball. In fantasy football there is a lot of debate around if a fantasy manager should take a running back or wide receiver in round 1 to achieve the highest chance of winning. This question continues throughout the draft, on researchers looking for optimal times to take certain positions based off player skill, player limits, and draft capital redundancy. The same could potentially be true in NBA where a team should be paying the best forward more than the best center or guard, to use more general, standard player positions, if a team wants to win the championship. This is what looking at salary allocation to the player classification attempted to achieve.

Another factor to look at was how a team could build a contender utilizing age. Are once good players now too old? Does the team need players on rookie contracts to compete in the NBA like other studies have shown are almost needed in the NFL? Does the best player on the team need to be close to the prime of their career? By clustering teams based off age, this was something that can looked at more closely. This was again looking at age allocation in the player classifications, similarly to the salary allocation.

A third factor to look at was how good coaches are. This is because some coaches will get more out of some players, while others might get more out of other players. Coach Thibodeau, currently of the New York Knicks, has a reputation of overachieving in the regular season due to playing his best players long minutes in the regular season. However, this same extended use often leads to more injuries late in the season causing his teams to underperform their regular season in the playoffs. Other coaches are better at developing young players, but not using their old players effectively, and vice versa. So, the goal was to look at coach and player classification to see if certain coaches outperform certain classifications or seasons in general.

The last factor that was looked at is whether the player was on the team last year or not. Continuality is often cited as a reason for the team doing better in a second year together, because they know how to “jell” with each other.

This whole study is unique in that there are multitude of studies trying to figure out how good a player will be when drafted, and multitude of studies and readymade statistics looking at how good a team is based on the collection of team statistics. However, from literature review, there could not be found much on how individual parts can have more hidden effects on the team. This is what looking at each of these individual factors will be taking a closer look at.

Data Source:

Almost all of the data sources were scraped from basketball-reference.com using Python. Most of the salary data came from basketball-reference.com, but some did come from hoopshype.com also. The data was scraped to get individual player’s and team’s statistics going back to 1964 for draft information, 1985 for salary information, and 1980 for all other information. The 1980 season is the first that had all the information that was needed outside of salary which was 1985, and the first draft class in 1980 was the 1964 draft class. The data used will be game total statistics, per 100 statistics, adjusted shooting statistics, advanced statistics, and general information on the players and teams in age, salary, coach, and most importantly of all as it was the target variable team playoff success.

For those who want to take the code that is attached to this in the zip file, scraping takes a long time to run so beware of that. Also, for pity points to the graders I had never scraped before so this probably took the most time in this research in over 80 hours by itself.

Methodology:

The classification of players was done using k-Means clustering. This allowed the classification to be unsupervised. Items that were be included into this unsupervised model were the season long box score statistics (games, games started, minutes played, field goals made, field goals attempted, offensive rebounds, steals, turnovers, points, etc.), per 100 offensive and defensive player ratings (which shows how many points the player's team scored and the other team scored while the player is on the court per 100 possessions), advanced statistics (PER, True Shooting %, 3 point attempt rate, offensive rebound percentage, TOV%, USG%, WS, WS/48 minutes, BPM, etc.), age, years played, salary, draft pick, and previous playoff success (an array of playoff appearances, first round of playoff victories, conference semifinal victories, NBA championship appearances, and NBA championship victories up to that year and including that year), team success in regular season wins/ losses, team offensive and defensive ratings, etc.

The team factor analysis was done with K-Nearest Neighbors classifiers (K-NN). This is because a team wide array made up with factors attached with the individual player performance from the season before was determined to the hypothesized best way to test full teams and their playoff success. The goal was to group them with each other if they were successful and group those not successful together also. Therefore, using other teams as neighbors seemed like a logical choice.

So, for example the salary model for a team looked in theory like…

[[Team], [Year], [Playoff_Success], [Previous_Year_Cluster_forPlayer1, …, Previous_Year_Cluster_forPlayerN], [salary_forPlayer1, …, salary_forPlayerN],

There were several issues with the data that had to be worked around. The biggest issue for this was how to deal with newly added rookies or players who have missed time due to injury. Players who have missed the entire season had the last season available to them as their past performance. In other words, players who missed a season or more due to injury or some other reason (Michael Jordan going to play baseball) had the last season's cluster that was available attached to them regardless of how little they played or how long before it was. Rookies were an entirely different story since there was no past performance to pull upon. Instead, the basics were brought into the model and weighted against all rookies at their draft pick throughout the whole model's running, and the mode cluster of those was chosen. This is to say the analysis was predicting if the team added a generic rookie with similar age, height, weight, and draft position for year one what their typical cluster would be like. The first of these were added so that the type of player expected could be based off measurables, while draft position was added because it has been shown by multiple other studies to be one of the biggest predictors of success in rookies. Some other issues that were attempted to be worked around is the lack of salary data for some players even after 1985. These players had their salary adjusted to be the minimum of the other players on their team. Age was also missing for some players, and so the mean age was chosen in that case.

Another issue that might have impacted the analysis is that the encoding through scraping added characters that might make player continuality not accurate. There was attempts to go back and fix all of the encoding issues, and work around the ones that were not fixed, but there is no guarantee that it was done 100% accurately.

Additionally, some other data changes that had to be made were that the usage of salary was instead salary proportion in the team analysis. This was needed as the salary cap moves each year, and the lack of proportion would cause generational bias if it was done on the pure salary number. This also feature scales the data which is needed in K-NN. Age, player continuality would be done the same way as player performance. For player continuality the players were assigned “On the team last year”, “Rookie”, or “Not on the team last year” to test whether or not this helped the clustering predict playoff success in the coming year. These were made into factors for the actual team analysis. The same methodology was done with coaching, but the whole team would got the name(s) of the coaches for that year. The coaches were also listed as factors.

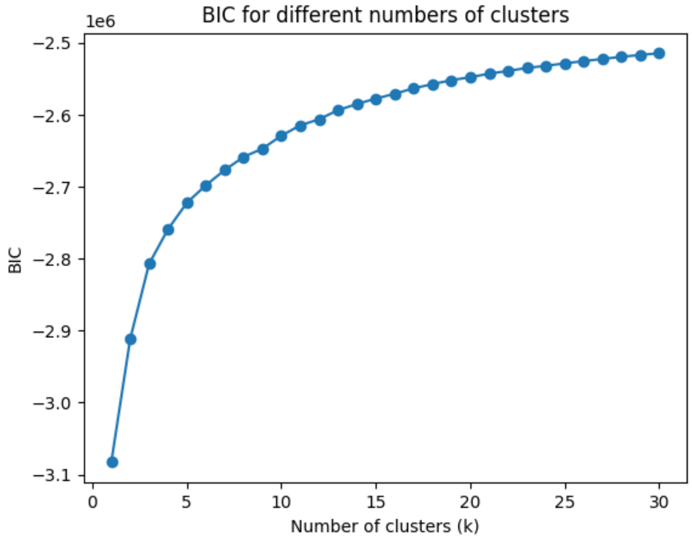
The target of “Playoff Success” was determined to be on a point score that was made up. Essentially, the further you went in the playoffs the more points the team got in a Fibonacci sequence. So not making the playoffs was 0, first round exit was 1, conference semifinals exit was 3, conference finals exit was 6, losing in the finals was 10, and winning the championship was 15 points. The weighted by distance was chosen for the K-NN in order to better reward the program for finding the best teams rather than just assigning the teams the typical 0 or 1 which is what most of the database is.

To test whether these models are successful, a random 20% of the data was held as test data. This test data was then used to test and predict how well the team would do the coming year. The prediction was be taken in the measure of accurate predictions (correct predictions divided by total). Additionally, for the base model and the addition of coaches feature importance was pulled to see which ones most impacted the mean decrease in accuracy.

Evaluation and Final Results:

To start for the clustering, an optimal k needed to be found. This was decided to be 20 as this is were the Bayesian information criterion slowed down. The optimal k was not chosen as this kept getting to be a higher and higher number, and this would start making less sense for being able to accurately categorize players more generally going forward. So the compromise was made for when the BIC plot slowed down (Figure 1).

Figure 1: BIC for different number of clusters



After the k was chosen for the K-Means Clustering was done.

Figure 2: Heatmap of Cluster Centers

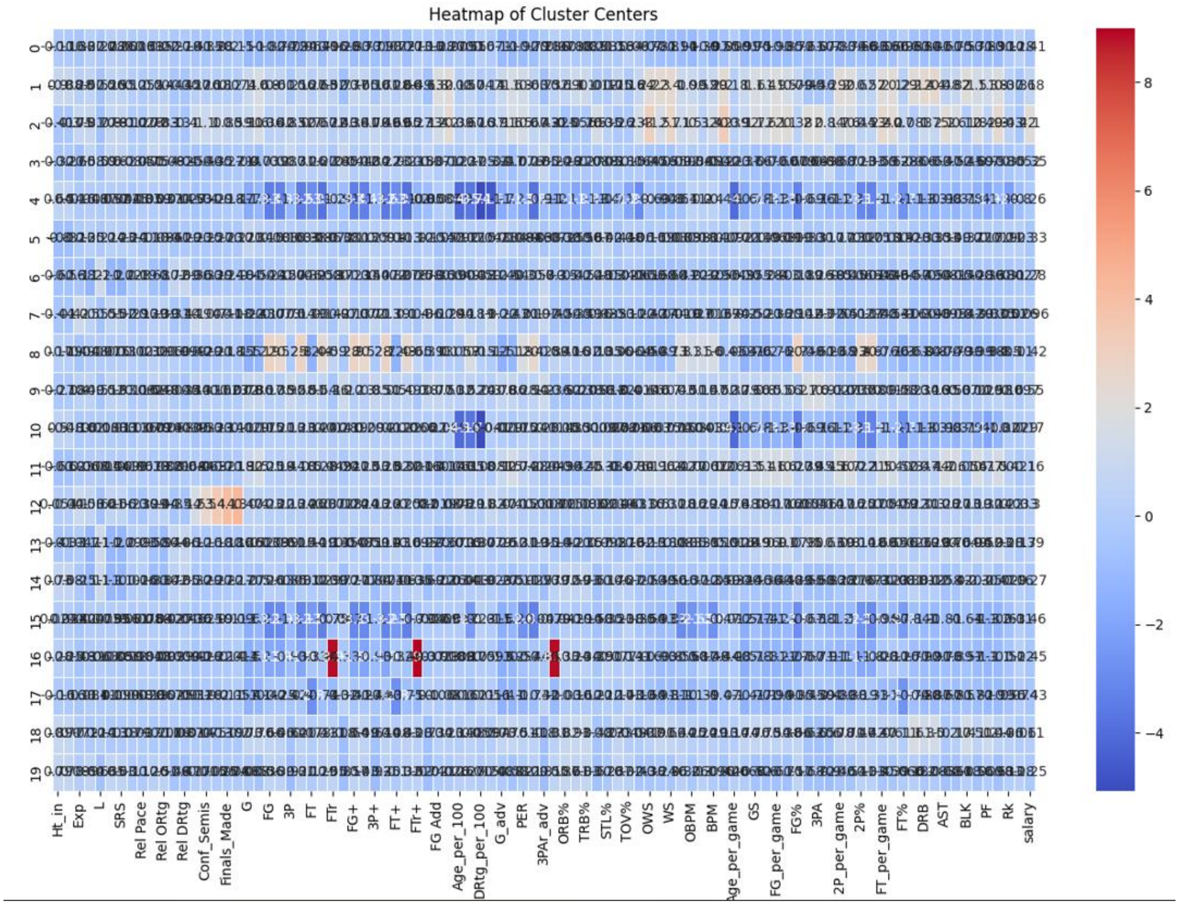
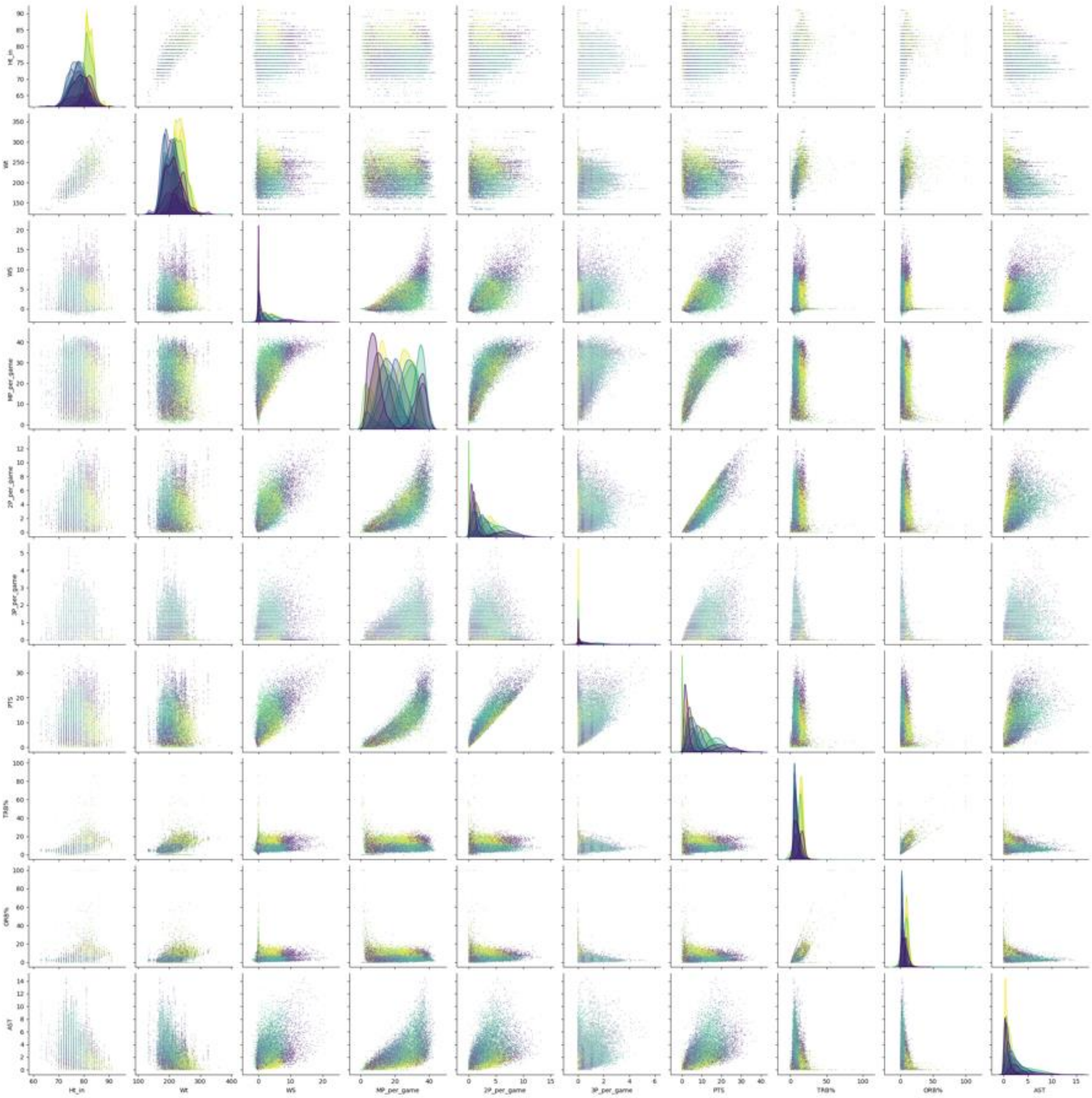


Figure 3: Some Selected Features Pair Plot



From this result, players now had a cluster assigned to them. Some key take aways from these results are that there was a player type hat seemed to excel at getting their team far into the playoffs in Cluster_12. This player did not seem to have their map standout in any other factor though. Some players with this given cluster were Draymond Green, Andre Iguodala, Danny Green, and Jae Crowder.

Additionally, top rookies and young stars seemed to often be assigned Cluster_1 if they were considered a big or Cluster_2 if they were more ball dominant. Some players with these given clusters not in the below table were Chet Holmgren, Domantas Sabonis, Victor Wembanyama, and Zion Williamson for Cluster_1 and De'Aaron Fox, Devin Booker, Jalen Brunson, and Giannis Antetokounmpo for Cluster_2. These clusters lit up the statistics scraped in the advanced section the most. Funnily enough it seemed like name might have been a chosen feature, and Cluster_4 got all of the European players. Otherwise, they play very similarly.

Table 1: Most Recent Three Players for Each Cluster (Sorted by Name and has a Salary)

Note that if a player was traded they might appear twice based off of their role and performance with the new team.

Year	Team	Player	Pos	Ht_in	Wt	Age	Exp	salary	Cluster
2024	ATL	AJ Griffin	SF	78	222	20	1	3712920	0
2024	OKC	Aleksej Pokusevski	PF	84	190	22	3	5574809	0
2024	LAL	Alex Fudge	SF	80	200	20	0	376405	0
2024	HOU	Alperen Sengun	C	83	235	21	2	3536280	1
2024	LAL	Anthony Davis	C	82	253	30	11	40600080	1
2024	MIA	Bam Adebayo	C	81	255	26	6	32600060	1
2024	MIN	Anthony Edwards	SG	76	225	22	3	13534817	2
2024	NOP	CJ McCollum	PG	75	190	32	10	35802469	2
2024	MIL	Damian Lillard	PG	74	195	33	11	45640084	2
2024	DAL	Alex Fudge	SF	80	200	20	0	376405	3
2024	ORL	Anthony Black	PG	79	198	20	0	7245480	3
2024	IND	Ben Sheppard	SG	78	190	22	0	2537172	3
2024	HOU	Boban Marjanovi?	C	88	290		8	2891467	4
2024	ATL	Bogdan Bogdanovi?	SG	77	220		6	18700000	4
2024	DET	Bojan Bogdanovi?	PF	79	226		9	20000000	4
2024	ORL	Cole Anthony	PG	74	185	23	3	5539771	5
2024	MIN	Kyle Anderson	PF	81	230	30	9	9219512	5
2024	ORL	Markelle Fultz	PG	76	209	25	6	17000000	5
2024	CHO	Aleksej Pokusevski	PF	84	190	22	3	5574809	6
2024	CHO	Amari Bailey	PG	77	185	19	0	559782	6
2024	WAS	Anthony Gill	PF	79	230	31	3	1997238	6
2024	HOU	Aaron Holiday	PG	72	185	27	5	2346614	7
2024	NYK	Alec Burks	SG	78	214	32	12	10489600	7
2024	MIL	Cameron Payne	PG	75	183	29	8	9391467	7
2024	SAS	Charles Bassey	C	82	235	23	2	2600000	8
2024	DAL	Daniel Gafford	PF	82	234	25	4	12402000	8
2024	BOS	Drew Peterson	PF	81	205	24	0	395708	8
2024	DEN	Aaron Gordon	PF	80	235	28	9	21266182	9
2024	IND	Aaron Nesmith	SF	77	215	24	3	5634257	9
2024	OKC	Aaron Wiggins	SG	78	200	25	2	1836096	9
2024	CHO	Davis Bertans	PF	82	225	31	7	17000000	10
2024	MEM	Maozinha Pereira	SF	80	177	23	0	128685	10
2024	CHO	Theo Maledon	PG	76	175	22	3	559782	10

2024	NOP	Brandon Ingram	SF	80	190	26	7	33833400	11
2024	UTA	Collin Sexton	SG	75	190	25	5	17525000	11
2024	ATL	Dejounte Murray	SG	77	180	27	6	18214000	11
2024	GSW	Draymond Green	PF	78	230	33	11	22321429	12
2024	DEN	Kentavious Caldwell-Pope	SG	77	204	30	10	14704938	12
2024	GSW	Kevon Looney	C	81	222	27	8	7500000	12
2024	DET	Alec Burks	SG	78	214	32	12	10489600	13
2024	POR	Anfernee Simons	SG	75	181	24	5	24107143	13
2024	DET	Ausar Thompson	SF	79	215	21	0	7977420	13
2024	CHI	Adama Sanogo	PF	81	245	21	0	559782	14
2024	MEM	Bismack Biyombo	C	80	255	31	12	6194075	14
2024	MEM	Brandon Clarke	PF	80	215	27	4	12500000	14
2024	OKC	Adam Flagler	SG	75	185	24	0	205897	15
2024	CHI	Andrew Funk	SG	77	200	24	0	160857	15
2024	IND	Daniel Theis	C	80	245	31	6	9108387	15
2024	TOR	Markquis Nowell	SG	68	160	24	0	559782	16
2024	IND	Quenton Jackson	PG	77	175	25	1	135120	16
2022	SAC	Emmanuel Mudiay	PG	75	200	25	5	111457	16
2024	ORL	Admiral Schofield	PF	77	241	26	3	559782	17
2024	NYK	DaQuan Jeffries	SG	77	230	26	4	1341538	17
2024	CLE	Emoni Bates	SF	82	170	20	0	559782	17
2024	HOU	Amen Thompson	SF	79	209	21	0	8809284	18
2024	CHI	Andre Drummond	C	83	279	30	11	3360000	18
2024	BRK	Ben Simmons	PG	82	240	27	5	37893408	18
2024	SAC	Alex Len	C	84	250	30	10	3196448	19
2024	OKC	Bismack Biyombo	C	80	255	31	12	6194075	19
2024	ATL	Bruno Fernando	C	81	240	25	4	2581522	19

With the clustering complete, the analysis for the K-Nearest Neighbors for team wide success was ready to be analyzed. Table 2 is an example of what some of the 2024 season teams looked like. While Table 3 is each models accuracy with the predictions.

Table 2: Part of First Six Teams in Alphabetic Order

Team	Year	Playoff_Success	Previous_Year_Cluster
ATL	2024	0	[3.0, 4.0, 14.0, 18.0, 9.0, 11.0, 8.0, 6.0, 3.0, 3.0, 19.0, 18.0, 7.0, 13.0, 0.0, 11.0, 0.0, 3.0,
BOS	2024	15	[9.0, 3.0, 9.0, 18.0, 3.0, 2.0, 2.0, 3.0, 0.0, 2.0, 4.0, 4.0, 3.0, 19.0, 19.0, 6.0, 3.0, 9.0, 3.0, 1
BRK	2024	0	[6.0, 18.0, 3.0, 9.0, 3.0, 19.0, 10.0, 13.0, 9.0, 3.0, 17.0, 0.0, 13.0, 6.0, 9.0, 9.0, 1.0, 14.0, 9
CHI	2024	0	[18.0, 9.0, 19.0, 18.0, 5.0, 9.0, 3.0, 11.0, 18.0, 3.0, 9.0, 0.0, 4.0, 18.0, 9.0, 3.0, 12.0, 2.0]
CHO	2024	0	[3.0, 6.0, 13.0, 6.0, 6.0, 10.0, 3.0, 13.0, 9.0, 0.0, 6.0, 6.0, 13.0, 18.0, 18.0, 16.0, 11.0, 18.
CLE	2024	3	[9.0, 18.0, 12.0, 11.0, 3.0, 2.0, 17.0, 1.0, 9.0, 9.0, 3.0, 1.0, 9.0, 9.0, 18.0, 3.0, 4.0, 3.0]

Table 3: Each Models Accuracy

Base KNN model

Accuracy: 0.44534412955465585

Classification Report:

```
         precision   recall  f1-score   support


      0     0.49     0.83     0.62      109

      1     0.33     0.21     0.26       75

      3     0.38     0.09     0.15       33

      6     0.00     0.00     0.00       12

     10      0.00     0.00     0.00        9

     15      0.00     0.00     0.00        9


  accuracy                    0.45      247

 macro avg      0.20     0.19     0.17      247

weighted avg      0.37     0.45     0.37      247
```

Salary

Accuracy: 0.4692982456140351

Classification Report:

```
         precision   recall  f1-score   support


      0     0.54     0.85     0.66      119

      1     0.14     0.09     0.11       56

      3     0.20     0.04     0.06       27

      6     0.00     0.00     0.00       12

     10      0.00     0.00     0.00        9
```

```
      15       0.00      0.00      0.00         5


   accuracy                         0.47       228

  macro avg       0.15      0.16      0.14       228

weighted avg       0.34      0.47      0.38       228
```

Age

```
Accuracy: 0.4493927125506073

Classification Report:

           precision    recall  f1-score   support


       0       0.52      0.89      0.65       109

       1       0.29      0.16      0.21        75

       3       0.10      0.03      0.05        33

       6       0.25      0.08      0.12        12

      10       0.00      0.00      0.00         9

      15       0.00      0.00      0.00         9


   accuracy                         0.45       247

  macro avg       0.19      0.19      0.17       247

weighted avg       0.34      0.45      0.36       247
```

Continuality

```
Accuracy: 0.43724696356275305

Classification Report:

           precision    recall  f1-score   support


       0       0.47      0.83      0.60       109
       1       0.35      0.24      0.29        75
       3       0.00      0.00      0.00        33
       6       0.00      0.00      0.00        12
      10       0.00      0.00      0.00         9
      15       0.00      0.00      0.00         9

    accuracy                           0.44       247
   macro avg       0.14      0.18      0.15       247
weighted avg       0.32      0.44      0.35       247
```

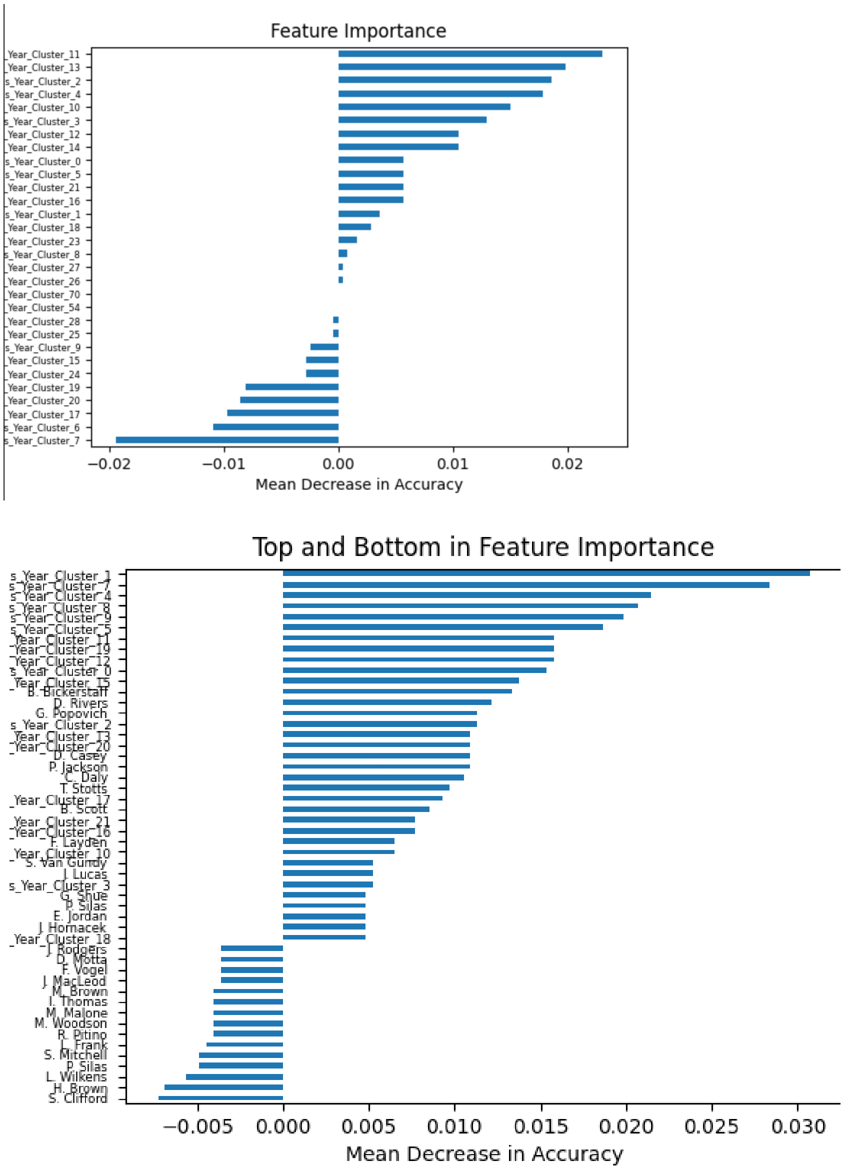
Coaches

```
Accuracy: 0.4493927125506073
Classification Report:
              precision    recall  f1-score   support

       0       0.55      0.67      0.60       109
       1       0.40      0.40      0.40        75
       3       0.21      0.21      0.21        33
       6       0.00      0.00      0.00        12
      10       0.00      0.00      0.00         9
      15       0.33      0.11      0.17         9

    accuracy                           0.45       247
   macro avg       0.25      0.23      0.23       247
weighted avg       0.40      0.45      0.42       247
```

From this it is shown that adding Coaches and Age to the model of Clusters helps improve the accuracy a little, but not much. While continuality actually decreased the accuracy. However, the one that showed the most promise of maybe improving the prediction was salary proportion allocation. While it was difficult to get behind the black box of what this means, the most obvious guess is paying bad players high contracts leads to more losing. This is predicted because it guesses only 5 champions when there are 0 while the others guessed 9.

Something that was done in an attempt to look at this further was by pulling out Feature Importance of the K-NN models.

<u>Figure 4: Feature Importance of Base Model (Top) and Coaches Model (Bottom)</u>



The top features in the base model are Cluster_11, Cluster_13, and Cluster_2. None of these clusters appeared to be bad players on high contacts to me. In fact, Cluster_2 has been hypothesized to be young, ball dominant stars. This shows that these features are not just helping choose missed playoffs, but also potential playoff success too. The coaches was done with the theory of trying to see which coaches had wild seasons (those that have negative Mean Decrease in Accuracy) so they are maybe more dependent on the players around them, and those that seem to always influence the model to pick more successful or not. Coaches like Doc Rivers and Greg Popovich are the first two coaches seen, showing that they influence their teams strongly.

Conclusion

What was seen in this research is that clustering players into 20 different types can give some insight into what type of player they are and potential successful predictions of playoff runs. While the K-NN was probably not the best choice of model to predict playoff teams there was still some value in it. Additionally, it appears organizing the teams by their player clustering provides the most insight into the coming year’s performance. However, adding salary proportion can also help with the prediction.

For future research I would like to look at other models to see if they can give better predictions with weighting and also more of percentage based guess for each grouping. Something like taking the same data and testing it with random forests. In particular, looking at salary data and trying to see how the proportion of salary increases the chance of guessing more accurately would be a wonderful next step.

Appendix

See the zip file for more graphs and full tables!