Contents

Introduction

In the basketball scene, there is a lot of professional basketball league that consist of many talented individual. However, each league differs in term of competition and playstyle. The two major basketball league would be The NBA and The EuroLeague. This research would use PCA and Cluster Analysis to differentiate the type of players that is in their respective league. The per-game stats would be used to create factors/cluster and differentiate the player type.

On top of that, a linear regression would be implemented to predict the factor that would increase/decrease an NBA player salary. The linear regression will use the previous season data and correlates it with the current season salary.

Data Acquisition

All datasets were web scraped using Beautiful Soup which is an HTML web scraper in Python. The initial coding was inspired by a GitHub page by (Cheema 2020) which scrapes the entire NBA Page for features/variables. However, the implementation of Beautiful Soup by (Sanchez 2019) is easier and thus both coding were used to scrape all three datasets. Three datasets were used and it is the NBA 2020-2021 salary, NBA players per-game stats and EuroLeague players per-game stats.

The Salary Dataset was scraped from basketball-references.com. The NBA per-game stats was scraped from NBA.com. The EuroLeague per-game stats was scraped from euroleague.net. All datasets variables description can be found on Appendix A.

Statistical Modelling

In this section, a linear regression model would be created and the dataset used was the 2019-2020 NBA player stats and the 2020-2021 NBA Player Salary. The concept behind this modelling is to use the previous year stats for each player and predict the salary for the upcoming season for NBA Players. Which category of stats would increase/decrease the salary of NBA Players.

Correlation Analysis

Prior to creating a regression model, a correlation analysis was produced in order to check the linear relationship between player stats and salary. Each variable should undergo a distribution analysis to decide what type of correlation analysis should be used on the datasets. However, the salary datasets show that it is not a normally distributed as it rejects the null hypothesis with a P-value less than 0.05 (Figure 1), thus allowing the usage of spearman correlation analysis.

Fitted Normal Distribution for seson20_21

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.2097243	Pr > D	<0.010
Cramer-von Mises	W-Sq	5.3953199	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	30.6127835	Pr > A-Sq	<0.005

Figure 1: Distribution Analysis for NBA Salary 2020-2021

Spearman Correlation Coefficients, N = 356
Prob > |r| under H0: Rho=0

	Age	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%
seson20_21	0.39423 <.0001	0.33282 <.0001	0.57470 <.0001	0.70923 <.0001	0.67797 <.0001	0.68090 <.0001	0.11820 0.0257	0.44984 <.0001	0.47298 <.0001	0.12224 0.0211	0.60807 <.0001	0.62353 <.0001	-0.01962 0.7122

Spearman Correlation Coefficients, N = 356
Prob > |r| under H0: Rho=0

	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
seson20_21	0.06423 0.2267	0.59730 <.0001	0.59462 <.0001	0.13940 0.0084	0.36619 <.0001	0.62540 <.0001	0.58765 <.0001	0.61388 <.0001	0.54877 <.0001	0.33688 <.0001	0.65177 <.0001	0.41944 <.0001	0.68159 <.0001

Figure 2: Correlation Analysis between NBA 2019-2020 per-game stats and NBA 2020-2021 Salary

In order for a more accurate model, a high degree of linear relationship was fixed at +0.6 correlation coefficients with an addition of a P-value less than 0.05. A low P-value in this analysis would means that there is a lower chance of the results to happens by chance. From Figure 2, it can be seen that the appropriate variables for the modelling would be:

Variables:

GS, MP, FG, FGA, 2P, 2PA, FT, FTA, DRB, TRB, AST, STL, TOV, PTS.

Linear Regression

The initial regression model consists of all the relevant variables from the correlation analysis. However, the analysis from Figure 3 a) promotes poor modelling as there is a low adjusted R^2 and a lot of high P-value which indicates that the variables has zero effect to the modelling. Further Analysis was done and the final regression model would be on Figure 3 b).

Number of Observations Read	356
Number of Observations Used	356

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1.936645E16	1.383318E15	36.79	<.0001
Error	341	1.282265E16	3.760309E13		
Corrected Total	355	3.21891E16			

Root MSE	6132136	R-Square	0.6016
Dependent Mean	8858324	Adj R-Sq	0.5853
Coeff Var	69.22456		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1761203	1251675	-1.41	0.1603
GS	1	32437	21823	1.49	0.1381
MP	1	-426236	122869	-3.47	0.0006
FG	1	7354200	10033512	0.73	0.4641
FGA	1	2851457	786267	3.63	0.0003
2P	1	717225	3948674	0.18	0.8560
2PA	1	-2585791	1017113	-2.54	0.0115
FT	1	6216339	3624225	1.72	0.0872
FTA	1	-974368	1411318	-0.69	0.4904
DRB	1	1815348	1188769	1.53	0.1277
TRB	1	87679	928285	0.09	0.9248
AST	1	1660687	380894	4.36	<.0001
STL	1	3105985	1202537	2.58	0.0102
TOV	1	-2995095	1237705	-2.42	0.0160
PTS	1	-3593564	3321687	-1.08	0.2801

Number of Observations Read	356
Number of Observations Used	356

Note: No intercept in model. R-Square is redefined.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	4.623063E16	6.604376E15	165.90	<.0001
Error	349	1.389376E16	3.981019E13		
Uncorrected Total	356	6.012439E16			

Root MSE	6309532	R-Square	0.7689
Dependent Mean	8858324	Adj R-Sq	0.7643
Coeff Var	71.22715		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
MP	1	-372028	70337	-5.29	<.0001
TRB	1	1116244	221180	5.05	<.0001
AST	1	1808778	367929	4.92	<.0001
STL	1	3036560	1218012	2.49	0.0131
TOV	1	-2611916	1154171	-2.26	0.0242
PTS	1	2227461	354901	6.28	<.0001
FG	1	-3430946	990650	-3.46	0.0006

Figure 3: a) Initial Regression Model b) Final Regression Model

Figure 3 b) shows that the Intercept was removed as a case where a zero independent variables would be an outlier. That means that an NBA Player would have to be injured for the whole season to log in 0 minutes on the court and that would not be an appropriate modelling as there is no variables on injury susceptibility in the acquired datasets.

The final model shows greater Adjusted R^2 which shows that the variables in that model has greater input toward the modelling. All variables would have low P-value which indicates that the variables do correlate with the model and proves that the model is a good-fit model.

The regression model indicates that a player would increase their next season salary if they have higher stats in PTS, AST, STL, TRB while having higher stats in TOV, FG, MP would decrease the salary. This model aligns with the current scenario in the NBA where a player who can create the most point while limiting their field goals and achieve it in a lower minutes played proves to be a very efficient player. Having additional stats in AST, STL and TRB also helps players to gain more salary. The parameter estimates indicate how much of an increase/decrease in salary would players gain if they increase the stated variables by a unit.

Statistical Analysis (Comparing Player-Type in NBA & EuroLeague)

There are two major sports-league in the basketball scene which are NBA and Euroleague. Despite playing the same sports, both league have different playstyles and attract different niche of basketball-fans. These two analysis would try to differentiate the leagues by using their players per-game stats. The variables used would be the same for each dataset.

Principal Component Analysis, PCA

PCA is a reduction method that aims to reduce a larger set of variables into smaller variables. PCA would then give factor patterns that would then be analyse and named. Basketball players have different playstyles and PCA would use the per-game stats to categorize them.

Assumptions:

1. Multiple continuous variables (both per-game stats for each league are continuous)
2. Linear relationship between variables
3. Sample size, >150 (assumptions met)
4. No significant outliers (This analysis assumed that there are no significant outliers that would defect the factor patterns)

The second assumptions would need to undergo correlation analysis. However, since there are too many variables to be tested among each other, only 5 variables would be checked for linearity on both datasets. Figure 4 shows that both datasets have linear relationship between variables.

Spearman Correlation Coefficients, N = 484
Prob > |r| under H0: Rho=0

	MP	3P%	FT%	STL	PTS
MP	1.00000	0.29854 <.0001	0.22444 <.0001	0.74592 <.0001	0.92796 <.0001
3P%	0.29854 <.0001	1.00000	0.27389 <.0001	0.14573 0.0013	0.35505 <.0001
FT%	0.22444 <.0001	0.27389 <.0001	1.00000	0.10272 0.0238	0.26904 <.0001
STL	0.74592 <.0001	0.14573 0.0013	0.10272 0.0238	1.00000	0.65580 <.0001
PTS	0.92796 <.0001	0.35505 <.0001	0.26904 <.0001	0.65580 <.0001	1.00000

Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	MP	3P%	FT%	STL	PTS
MP	1.00000 300	0.18390 0.0029 261	0.20334 0.0009 263	0.73488 <.0001 300	0.90221 <.0001 300
3P%	0.18390 0.0029 261	1.00000 261	0.21349 0.0008 242	0.14974 0.0155 261	0.28662 <.0001 261
FT%	0.20334 0.0009 263	0.21349 0.0008 242	1.00000 263	0.20664 0.0007 263	0.21663 0.0004 263
STL	0.73488 <.0001 300	0.14974 0.0155 261	0.20664 0.0007 263	1.00000 300	0.68391 <.0001 300
PTS	0.90221 <.0001 300	0.28662 <.0001 261	0.21663 0.0004 263	0.68391 <.0001 300	1.00000 300

Figure 4: a) Spearman Coefficients – NBA b) Spearman Coefficients - EuroLeague

Prior to analysing both factor patterns, a rotational method specifically orthogonal rotation (varimax) would be used on both datasets. This is to easily interpret the new artificial variables as varimax would maximized the sum of variances of the squared loadings as all factor patterns would have either large number or close to zero.

The number of factors to retain depends on the Eigenvalues of the Correlation Matrix and the Scree Plot. The Eigenvalues would need to be close to one and the decided min point would be +0.8 instead of 1 as this is to increase the explained variance min point which was decided at 0.9. This is to create a more accurate factor patterns that would explain player types more accurately. There would be 24 variables used for both datasets.

PCA – NBA

Number of Factors

Eigenvalues of the Correlation Matrix: Total = 24 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	12.2048202	8.5224271	0.5085	0.5085
2	3.6823931	1.5861884	0.1534	0.6620
3	2.0962048	0.8464304	0.0873	0.7493
4	1.2497744	0.3482063	0.0521	0.8014
5	0.9015680	0.1313626	0.0376	0.8389
6	0.7702054	0.0876801	0.0321	0.8710
7	0.6825253	0.1639676	0.0284	0.8995

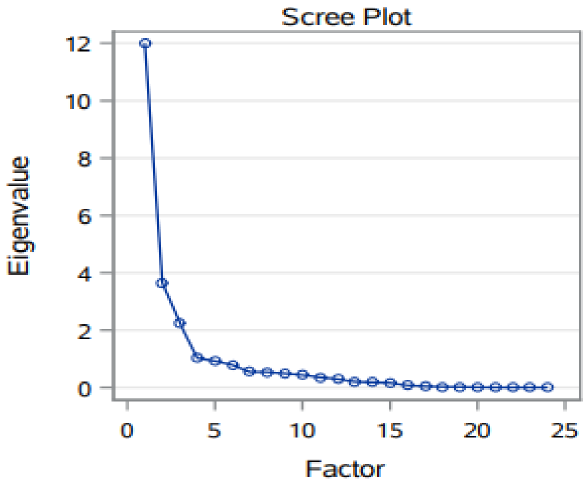

Figure 5: a) NBA Eigenvalues b) NBA Scree Plot

Figure 5 a) shows that the eigenvalue for the 7th factor does not meet the requirement for the eigenvalues of +0.8 while figure 5 b) is a scree plot that supports retaining 6 factors. Taking only 5 factors would suffice but adding one more factor would help explaining the dataset even clearer.

Rotational (Varimax)

Variance Explained by Each Factor					
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
12.204820	3.682393	2.096205	1.249774	0.901568	0.770205

Variance Explained by Each Factor					
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
8.7137535	4.0326182	2.8292280	2.2176054	1.9771917	1.1345691

Figure 6: a) Initial Variance Explained b) Rotated Variance Explained

Based on Figure 6, the application of the varimax further increase the variance explained by each factor except for the first factor. This indicates that the variance is now more evenly spread and allow for better understanding of the artificial variables.

Final Communality Estimates: Total = 20.904966								
G	MP	FG	FGA	FG%	3P	3PA	3P%	2P
0.67964408	0.92683596	0.96456201	0.96479652	0.96339567	0.89416205	0.86868261	0.78909456	0.93596271
2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB
0.92021403	0.84456953	0.95175745	0.88921888	0.88952668	0.92110562	0.87804398	0.88105487	0.92807438
AST	STL	BLK	TOV	PF	PTS			
0.78200120	0.73413578	0.70916767	0.88247448	0.72169840	0.98478681			

Figure 7: Final Communality for NBA

In PCA, the communality is the total variance for the datasets/variables. In figure 7, all variables after rotation have >0.5 communality which displays great correlation between one another when only using 6 factors. The final rotated communalities are short by 3 points from 24 which is the number of variables used but the factor retained was still 6.

Factor Pattern Analysis

Rotated Factor Pattern						
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
G	0.20377	0.24181	0.24807	0.65841	0.18595	0.22369
MP	0.65753	0.35303	0.03154	0.54483	0.24781	0.10305
FG	0.86503	0.30601	0.12661	0.21908	0.22185	0.09692
FGA	0.86953	0.22153	-0.01169	0.27222	0.26750	0.11762
FG%	0.11075	0.27811	0.92976	-0.04527	0.00028	-0.08539
3P	0.48599	-0.10744	-0.09168	0.42159	0.64070	0.22313
3PA	0.52679	-0.10587	-0.15980	0.43558	0.56292	0.21869
3P%	0.04632	0.00768	0.16485	0.00487	0.87157	-0.00729
2P	0.84145	0.42650	0.19955	0.07331	-0.02302	0.01731
2PA	0.87132	0.37851	0.08710	0.09774	0.00883	0.02306
2P%	0.02058	0.13491	0.89759	0.08367	-0.11288	-0.02325
eFG%	0.03991	0.16385	0.89269	0.09097	0.34360	-0.00870
FT	0.90656	0.18715	0.02619	0.03435	0.06519	0.16193
FTA	0.90480	0.24750	0.04932	0.02483	0.03151	0.07457
FT%	0.16762	-0.07510	-0.09466	0.07349	0.08408	0.93056
ORB	0.21032	0.84630	0.27387	-0.01362	-0.19525	-0.06539
DRB	0.49143	0.75434	0.11904	0.22501	0.07370	0.01710
TRB	0.43620	0.82466	0.17261	0.16711	-0.00102	-0.00468
AST	0.79969	-0.07741	-0.01546	0.36584	0.01133	-0.04792
STL	0.51844	0.15300	-0.04780	0.64825	-0.02427	-0.13728
BLK	0.11181	0.81953	0.15022	0.03154	0.02741	-0.02693
TOV	0.88104	0.18836	0.00436	0.25930	0.05438	-0.02352
PF	0.33786	0.60987	0.10829	0.45649	0.11959	-0.03455
PTS	0.88452	0.24646	0.08433	0.22089	0.25963	0.13548

Figure 8: NBA Factor Pattern

The cut-off point for a good factor loading is +0.4 and based on Figure 8, the highlighted variables for each pattern meet the cut-off point. Each factor would then be named as a player-type based on the highlighted variables.

The Components:

Factor 1: **High usage all-around player** that dominates in most per-game stats. This component indicates the superstars on each team.

Factor2: **Traditional Big player** who take a lot of rebound, blocks shot and finish inside the 3-point lines mainly using dunks or layup. This can also be seen on the high factor of number of Personal-Foul.

Factor 3: **Efficient Bench Player** who scores their point efficiently without taking too much shots and serve as probably the few last scoring options. They will however make their shots.

Factor 4: **3 and D Guard.** Athletic player (usually smaller player) who mainly used as a defender and mainly scores from 3 point shots. However, these players would not be efficient and thus would not be a high usage player and would just compliment the superstar. They would make the right passes to control the offense thus having a good amount of assist factor.

Factor 5: **3 Point Shooter** that shoots a lot of 3 points shots and make a lot of them.

Factor 6: **Bad Bench Player** as it has a lot of negative factor loadings. Only shoots free-throw efficiently due to the limited minutes and limited shot-attempt.

PCA-EuroLeague

Number of Factors

Eigenvalues of the Correlation Matrix: Total = 24 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	11.9879561	8.3505728	0.4995	0.4995
2	3.6373833	1.3912474	0.1516	0.6511
3	2.2461359	1.2181170	0.0936	0.7446
4	1.0280190	0.1018936	0.0428	0.7875
5	0.9261254	0.1406764	0.0386	0.8261
6	0.7854489	0.2398656	0.0327	0.8588
7	0.5455833	0.0048540	0.0227	0.8815

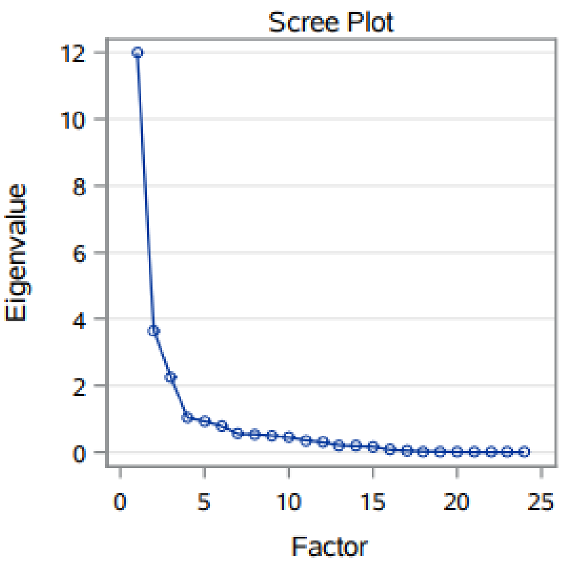

Figure 9: a) EuroLeague Eigenvalues b) EuroLeague Scree Plot

Figure 9 a) shows that the eigenvalue for the 7th factor does not meet the requirement for the eigenvalues of +0.8 while figure 9 b) is a scree plot that supports retaining 6 factors. Taking only 5 factors would suffice but adding one more factor would help explaining the dataset even clearer.

Rotational (Varimax)

Variance Explained by Each Factor					
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
11.987956	3.637383	2.246136	1.028019	0.926125	0.785449

Variance Explained by Each Factor					
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
5.5717242	3.9692095	3.5521315	3.1123577	2.6884816	1.7171641

Figure 10: a) Initial Variance Explained b) Rotated Variance Explained

Based on Figure 10, the application of the varimax further increase the variance explained by each factor except for the first factor. This indicates that the variance is now more evenly spread and allow for better understanding of the artificial variables.

Final Communality Estimates: Total = 20.611069

G	MP	FG	FGA	FG%	3P	3PA	3P%	2P
0.68046494	0.92028511	0.95743625	0.95489839	0.96946009	0.95242648	0.93376147	0.69570317	0.92428696

2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB
0.90143841	0.78475610	0.94959435	0.89122358	0.88802511	0.83740219	0.84935459	0.85495821	0.92429081

AST	STL	BLK	TOV	PF	PTS
0.81505379	0.66830271	0.69509277	0.84618813	0.73110952	0.98555552

Figure 11: Final Communality for EuroLeague

In Figure 11, all variables after rotation have >0.5 communality which displays great correlation between one another when only using 6 factors. The final rotated communalities are short by 3 points from 24 which is the number of variables used but the factor retained was still 6.

Factor Pattern Analysis

Rotated Factor Pattern						
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
G	0.20099	0.32097	0.31147	0.20993	0.11040	0.61950
MP	0.47731	0.37746	0.48133	0.16756	0.44701	0.30069
FG	0.69806	0.34819	0.42617	0.25269	0.31295	0.07414
FGA	0.67581	0.21797	0.53184	0.07273	0.38310	0.12557
FG%	0.16185	0.25298	0.00934	0.93089	0.05304	0.09900
3P	0.26894	-0.11991	0.89974	0.02795	0.19956	0.12487
3PA	0.29632	-0.12056	0.85612	-0.07180	0.27301	0.13705
3P%	-0.01638	-0.18687	0.66529	0.41904	0.00783	0.20556
2P	0.71299	0.50384	0.00091	0.29814	0.26982	0.01986
2PA	0.75413	0.42790	0.05772	0.16998	0.33556	0.06935
2P%	0.17045	0.23276	-0.01802	0.81582	0.09327	0.16414
eFG%	0.07362	0.14500	0.23995	0.92312	0.04181	0.10808
FT	0.86223	0.12239	0.21197	0.04651	0.12750	0.26355
FTA	0.86929	0.20125	0.14508	0.07725	0.14560	0.20893
FT%	0.29148	0.04732	0.18897	0.21431	0.20148	0.79244
ORB	0.33362	0.78771	-0.21343	0.18869	-0.07984	0.17331
DRB	0.42961	0.74552	0.15486	0.19001	0.15857	0.17136
TRB	0.42314	0.81480	0.02400	0.20210	0.07965	0.18327
AST	0.37687	-0.21427	0.21687	-0.02807	0.74712	0.14528
STL	0.18319	0.32931	0.41366	0.11679	0.57678	0.09418
BLK	0.02694	0.80454	-0.14132	0.15512	0.05179	-0.01932
TOV	0.53048	0.07197	0.19527	0.07608	0.71386	0.07805
PF	0.09663	0.38194	0.11803	0.19736	0.56543	0.45089
PTS	0.73167	0.25374	0.50059	0.19403	0.28273	0.13285

Figure 12: EuroLeague Factor Pattern

The cut-off point for a good factor loading is +0.4 and based on Figure 12, the highlighted variables for each pattern meet the cut-off point. Each factor would then be named as a player-type based on the highlighted variables.

The Components:

Factor 1: **All-around player** that would scores mainly inside the 3-point line with layups/dunks. Would also lead in assist and control the pace of the game. Mainly the most versatile player on the team.

Factor 2: **Traditional Big player** who take a lot of rebound, blocks shot and finish inside the 3-point lines mainly using dunks or layup. This can also be seen on the high factor of number of Personal-Foul.

Factor 3: **Two-Way Player**. These players would attempt and make a lot of 3 point shots with great efficiency while being their team best perimeter defenders with a lot of steals. They would attempt a lot of shots and would be among their team top scorers.

Factor 4: **Shooter**. Efficient player that would come off the bench to score standstills shots.

Factor 5: **Playmaker**. Know the strategy of the game very well and would create the most points for others while being a good defender. Would not be efficient scorer as most playmaker lacked in height which is a disadvantage to score.

Factor 6: **Benchwarmer**. Would complement the starters but would not lead in any category.

Cluster Analysis

A hierarchical clustering was done on both datasets and the agglomerative method would be applied. The number of cluster used was first tested by using the number of PCA which was 6 on both datasets. Further analysis would be done below to choose the appropriate number of clusters.

NBA

Cluster History								
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Tie
15	CL54	CL24	106	0.0035	.915	359	37.6	
14	CL16	CL32	15	0.0050	.910	364	7.5	
13	CL47	CL21	26	0.0050	.905	373	14.8	
12	CL23	CL26	93	0.0071	.898	376	63.0	
11	CL19	CL25	76	0.0074	.890	384	40.9	
10	CL12	CL74	96	0.0096	.881	389	51.0	
9	CL11	CL17	127	0.0113	.869	395	46.2	
8	CL10	CL22	123	0.0119	.857	409	43.7	
7	CL18	CL165	62	0.0142	.843	428	81.1	
6	CL8	CL15	229	0.0195	.824	447	76.2	
5	CL20	CL9	152	0.0353	.788	446	112	
4	CL14	CL13	41	0.0406	.748	475	58.7	
3	CL6	CL7	291	0.0637	.684	521	180	
2	CL5	CL4	193	0.1189	.565	627	151	
1	CL2	CL3	484	0.5653	.000	.	627	

Figure 13: NBA Cluster Analysis

Based on Figure 13, there are two significant increase in Pseudo t-Squared which is from 6-5 and 15-14. Both clusters would have a significant value of Pseudo F Statistic and an R-Square value close to one. The R-Square value for the higher amount of cluster would have the higher one but that is inevitable. Therefore, the smaller cluster at 6 was chosen to simplify the type of players in the NBA. Appendix B shows the first 6 players per-game stats for each cluster. From the clustering table, the player type was determined as below:

Cluster Name:

Cluster 1: **Bench-Player**

Cluster 2: **3-point Shooter**

Cluster 3: **3 & D Playmaker**. These players would be great in 3-point shooting, defence as proven by the steals number and have a good assist number.

Cluster 4: **Rebounding Big-Man**. Since this dataset would have the variables position, it can be seen that this cluster belongs to big man who block shots and grab multiple rebounds.

Cluster 5: **Superstar**. Taken from the high points, assist, minutes-played per-game average.

Cluster 6: **Scoring Big-Man** who grabs rebounds and score at rim.

EuroLeague

Cluster History								
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Tie
11	CL23	CL39	22	0.0075	.901	263	24.2	
10	CL14	CL18	109	0.0083	.893	269	41.4	
9	CL15	CL12	56	0.0095	.883	276	20.4	
8	CL26	CL17	15	0.0128	.871	281	21.7	
7	CL21	CL9	100	0.0162	.854	287	36.1	
6	CL11	CL30	24	0.0174	.837	302	25.3	
5	CL10	CL6	133	0.0311	.806	306	65.6	
4	CL7	CL8	115	0.0659	.740	281	92.6	
3	CL4	CL13	147	0.0900	.650	276	78.6	
2	CL5	CL64	153	0.1061	.544	356	171	
1	CL2	CL3	300	0.5440	.000	.	356	

Figure 15: Euroleague Cluster Analysis

Based on Figure 15, the Pseudo t-Squared shows that the appropriate amount of cluster would be only 6 as it is the first significant increase in value going down. The R-Square also justify the number of cluster at 6 as it is close to 1. Appendix C shows the first 6 players per-game stats for each cluster. From the clustering table, the player type was determined as below:

Cluster 1: **Bench-Player/Rookie**. Barely got any playing time.

Cluster 2 and Cluster 3: **Two-Way Player**. This two clusters have some similar stats which focuses on defences and 3-points shooting. However, cluster 3 takes more 3-pointer while cluster 2 took more 2 pointers but it is almost similar.

Cluster 4: **Efficient shooter**

Cluster 5: **All–Around Big-Man**. Players who scores mainly inside while grabbing multiple rebounds. Can facilitate by getting assists.

Cluster 6: **All-Around Scorer** who can assist the game but also scores the most points.

Discussion/Comparison

From the PCA and Cluster Analysis, it can be concluded that the NBA does differ a bit compare to the EuroLeague. The NBA is more individually-oriented as it usually depends on a player to score a lot of points while the EuroLeague is more on team-oriented playstyle as the scoring duties relies on the whole team. This justification can be seen during the cluster analysis where cluster 3 and cluster 2 propose a similar playstyle. On top of that, the minutes played factor pattern on the EuroLeague is evenly spread across 6 pattern while the NBA relies more on the Superstar.

Conclusion

In conclusion, The Linear Regression model indicates that a two-way player that is efficient on both offense and defence would highly likely get the most salary. However, further improvement could be done as the player celebrity status would greatly impact the regression model.

On top of that, PCA and Cluster Analysis proves that there is a difference between NBA and EuroLeague in terms of the type of player around the league. There are more team-oriented players in the EuroLeague while the NBA is more individually oriented with more high usage players. An improvement that could be done is to added more features such the rookie status, the veteran status or the usage rate for each players.

References

I. *Welcome to EUROLEAGUE BASKETBALL* [Online]. Available at: https://www.euroleague.net/ [Accessed: 11 January 2021].
II. *The Official of The NBA* [Online]. Available at: https://www.nba.com/ [Accessed: 11 January 2021].
III. *Basketball Reference* [Online]. Available at: https://www.basketball-reference.com/contracts/players.html/ [Accessed: 11 January 2021].
IV. Sanchez. O, *Web Scraping NBA Stats*[Online]. Available at: https://towardsdatascience.com/web-scraping-nba-stats-4b4f8c525994/ [Accessed: 11 January 2021].
V. Cheema. A, *Nba-Classification*[Online]. Available at: https://github.com/ahmed-cheema/nba-classification/blob/master/part_one/data_collection.py/ [Accessed: 11 January 2021]
VI.

Appendix

Appendix A:

Player: The name of the basketball player

TM: The name of the NBA team

Seson20_21: The salary earned for the NBA player

Pos: position played

Age: Age of player

G: Game Played

GS: Game Started

MP: Minute Played

FG: Field Goals Made

FGA: Field Goals Attempt

FG%: Field Goals Percentage

3P: 3-Pointer Made

3PA: 3-Pointer Attempt

3P%: 3-Pointer Percentage

2P: 2-Pointer Made

2PA: 2-Pointer Attempt

2P%: 2-Pointer Percentage

eFG%: Efficient Field Goals Percentage

FT: Free Throw Made

FTA: Free Throw Attempt

FT%: Free Throw Percentage

ORB: Offensive Rebound

DRB: Defensive Rebound

TRB: Total Rebound

AST: Assist

STL: Steal

BLOCK: Block

TOV: Turnover

PF: Personal Foul

PTS: Points

Appendix B:

Cluster 1:

Player	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Vlatko Čančar	PF	14	0	3.2	0.4	1.1	0.4	0.1	0.4	0.167	0.4	0.6	0.556	0.433	0.3	0.3	1	0.4	0.4	0.7	0.2	0.1	0.1	0.2	0.5	1.2	1
Charlie Brown	SG	10	0	4	0.6	1.9	0.316	0.3	0.9	0.333	0.3	1	0.3	0.395	0.5	0.5	1	0.2	0.2	0.4	0.2	0.2	0.2	0.3	0.3	2	1
Malik Newman	SG	1	0	4	0	2	0	0	1	0	0	1	0	0	2	2	1	0	0	0	0	0	0	1	0	2	1
Antonius Cleveland	SG	11	0	4.2	0.4	1.3	0.286	0	0.3	0	0.4	1	0.364	0.286	0.3	0.5	0.6	0.2	0.5	0.6	0.1	0.1	0.3	0.2	0.7	1	1
Zhaire Smith	SF	7	0	4.6	0.4	1.6	0.273	0	0.4	0	0.4	1.1	0.375	0.273	0.3	0.6	0.5	0	0.3	0.3	0.3	0.4	0	0.3	0.6	1.1	1
Dewan Hernandez	C	6	0	4.7	0.8	2.3	0.357	0.2	0.3	0.5	0.7	2	0.333	0.393	0.5	0.8	0.6	0.8	1.5	2.3	0.5	0.2	0	0.5	1	2.3	1

Cluster 2:

Player	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Anthony Tolliver	PF	55	13	15.9	1.2	3.3	0.357	0.9	2.6	0.338	0.3	0.7	0.432	0.492	0.4	0.5	0.724	0.6	2.2	2.8	0.8	0.3	0.2	0.6	1.3	3.6	2
Markieff Morris	PF	58	17	20.5	3.5	7.9	0.443	1.5	3.9	0.386	2	4	0.5	0.539	1.1	1.5	0.776	0.7	3.1	3.8	1.3	0.5	0.3	1.3	2.3	9.7	2
Allen Crabbe	SG	37	1	17.6	1.7	4.8	0.356	1	3.2	0.303	0.7	1.6	0.466	0.458	0.3	0.4	0.714	0.3	1.8	2.1	0.9	0.4	0.1	0.5	1.3	4.6	2
Gorgui Dieng	C	63	17	17.4	2.6	5.8	0.456	0.9	2.4	0.355	1.8	3.4	0.528	0.53	1.2	1.6	0.772	1.4	4.2	5.6	1.2	0.8	0.9	1	2	7.4	2
Mario Hezonja	PF	53	4	16.4	1.7	4.1	0.422	0.5	1.5	0.308	1.3	2.6	0.486	0.477	0.9	1.1	0.814	0.6	2.9	3.5	0.9	0.7	0.2	0.8	2	4.8	2
Frank Kaminsky	C	39	13	19.9	3.5	7.9	0.45	1.1	3.2	0.331	2.5	4.7	0.53	0.516	1.6	2.3	0.678	0.9	3.6	4.5	1.9	0.4	0.3	0.9	1.9	9.7	2

Cluster 3:

Player	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Malcolm Brogdon	PG	54	54	30.9	6	13.8	0.438	1.4	4.3	0.326	4.6	9.5	0.488	0.489	3.1	3.4	0.892	0.9	4	4.9	7.1	0.6	0.2	2.4	1.8	16.5	3
Mikal Bridges	SF	73	32	28	3.4	6.6	0.51	1	2.7	0.361	2.4	4	0.61	0.583	1.4	1.7	0.844	0.9	3.1	4	1.8	1.4	0.6	1	2.2	9.1	3
Eric Bledsoe	PG	61	61	27	5.5	11.5	0.475	1.2	3.5	0.344	4.3	8	0.533	0.528	2.8	3.5	0.79	0.7	3.9	4.6	5.4	0.9	0.4	2.4	2.1	14.9	3
OG Anunoby	SF	69	68	29.9	4.1	8.2	0.505	1.3	3.3	0.39	2.9	4.9	0.583	0.584	1	1.5	0.706	1.2	4.1	5.3	1.6	1.4	0.7	1.1	2.4	10.6	3
Goran Dragić	PG	59	3	28.2	5.4	12.3	0.441	2.1	5.7	0.367	3.3	6.5	0.505	0.526	3.2	4.2	0.776	0.5	2.7	3.2	5.1	0.7	0.2	2.4	2.1	16.2	3
Mike Conley	PG	47	41	29	4.9	12.1	0.409	2	5.4	0.375	2.9	6.6	0.437	0.494	2.4	3	0.827	0.7	2.5	3.2	4.4	0.8	0.1	2	2.2	14.4	3

Cluster 4:

Player	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Damian Jones	C	55	27	16.1	2.2	3.2	0.68	0	0.2	0.222	2.2	3.1	0.704	0.685	1.1	1.5	0.738	1.3	2.4	3.7	0.6	0.5	0.7	0.5	2.7	5.6	4
Jaxson Hayes	C	64	14	16.9	2.7	4	0.672	0	0.1	0.25	2.7	4	0.678	0.674	1.9	3	0.647	1.5	2.5	4	0.9	0.4	0.9	0.8	2.5	7.4	4
Dwight Howard	C	69	2	18.9	2.9	4	0.729	0	0.1	0.6	2.9	3.9	0.732	0.735	1.6	3.1	0.514	2.5	4.9	7.3	0.7	0.4	1.1	1.2	3.2	7.5	4
Ivica Zubac	C	72	70	18.4	3.3	5.3	0.613	0	0	0	3.3	5.3	0.616	0.613	1.7	2.3	0.747	2.7	4.8	7.5	1.1	0.2	0.9	0.8	2.3	8.3	4
Enes Kanter	C	58	7	16.9	3.4	6	0.572	0	0.1	0.143	3.4	5.9	0.581	0.573	1.2	1.7	0.707	2.8	4.6	7.4	1	0.4	0.7	1	1.7	8.1	4
Jonas Valančiūnas	C	70	70	26.4	6.2	10.6	0.585	0.5	1.3	0.352	5.7	9.3	0.617	0.606	2.1	2.9	0.74	3	8.2	11.3	1.9	0.4	1.1	1.8	2.6	14.9	4

Cluster 5:

Player	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
James Harden	SG	68	68	36.5	9.9	22.3	0.444	4.4	12.4	0.355	5.5	9.9	0.556	0.543	10.2	11.8	0.865	1	5.5	6.6	7.5	1.8	0.9	4.5	3.3	34.3	5
Jimmy Butler	SF	58	58	33.8	5.9	13.1	0.455	0.5	2.1	0.244	5.4	11	0.495	0.474	7.6	9.1	0.834	1.8	4.8	6.7	6	1.8	0.6	2.2	1.4	19.9	5
Russell Westbrook	PG	57	57	35.9	10.6	22.5	0.472	1	3.7	0.258	9.6	18.7	0.514	0.493	5.1	6.7	0.763	1.8	6.2	7.9	7	1.6	0.4	4.5	3.5	27.2	5
Kyle Lowry	PG	58	58	36.2	5.8	13.8	0.416	2.8	8	0.352	2.9	5.8	0.504	0.518	5.1	5.9	0.857	0.6	4.5	5	7.5	1.4	0.4	3.1	3.3	19.4	5
Stephen Curry	PG	5	5	27.8	6.6	16.4	0.402	2.4	9.8	0.245	4.2	6.6	0.636	0.476	5.2	5.2	1	0.8	4.4	5.2	6.6	1	0.4	3.2	2.2	20.8	5
LeBron James	PG	67	67	34.6	9.6	19.4	0.493	2.2	6.3	0.348	7.4	13.1	0.564	0.55	3.9	5.7	0.693	1	6.9	7.8	10.2	1.2	0.5	3.9	1.8	25.3	5

Cluster 6:

Player	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Hassan Whiteside	C	67	61	30	6.5	10.5	0.621	0.1	0.1	0.571	6.4	10.4	0.622	0.624	2.4	3.6	0.686	3.9	9.7	13.5	1.2	0.4	2.9	1.8	2.9	15.5	6
Andre Drummond	C	57	56	33	7.3	13.8	0.533	0.1	0.6	0.143	7.2	13.1	0.551	0.536	3	5.2	0.575	4.4	10.8	15.2	2.7	1.9	1.6	3.6	3.5	17.7	6
Giannis Antetokounmpo	PF	63	63	30.4	10.9	19.7	0.553	1.4	4.7	0.304	9.5	15	0.631	0.589	6.3	10	0.633	2.2	11.4	13.6	5.6	1	1	3.7	3.1	29.5	6
Karl-Anthony Towns	C	35	35	33.9	9	17.8	0.508	3.3	7.9	0.412	5.8	9.9	0.586	0.6	5.1	6.5	0.796	2.7	8.1	10.8	4.4	0.9	1.2	3.1	3.3	26.5	6
Ben Simmons	PG	57	57	35.4	6.6	11.4	0.58	0	0.1	0.286	6.5	11.2	0.583	0.581	3.2	5.2	0.621	2	5.8	7.8	8	2.1	0.6	3.5	3.3	16.4	6
Zion Williamson	PF	24	24	27.8	8.8	15	0.583	0.3	0.6	0.429	8.5	14.4	0.59	0.592	4.8	7.4	0.64	2.7	3.6	6.3	2.1	0.7	0.4	2.5	1.8	22.5	6

Appendix C:

Cluster 1:

Player	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Andrei Lopatin	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Nemanja Nenadic	1	3	0	1	0	0	0	0	0	1	0	0	1	2	0.5	0	0	0	0	0	0	1	1	1	1
Jimenez Millan	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Omer Can Ilyasoglu	1	3	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1
Jason George	1	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Ferrando Guillem	1	4	0	2	0	0	0	0	0	2	0	0	0	0	0	0	1	1	1	0	0	0	2	0	1

Cluster 2:

Player	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Kresimir Nikic	11	10.8	1.2	2.5	0.481	0	0.1	0	1.2	2.4	0.5	0.481	0.3	0.7	0.375	0.5	1.5	2	0.5	0.5	0.4	0.9	1.3	2.6	2
Kenneth Ogbe	18	8.6	0.9	2.2	0.436	0.3	0.7	0.385	0.7	1.4	0.462	0.5	0.3	0.3	0.833	0.2	1	1.2	0.3	0.3	0	0.6	0.8	2.4	2
Filip Covic	13	14.1	0.7	3.2	0.22	0.5	2.1	0.259	0.2	1.1	0.143	0.305	0.4	0.5	0.714	0.1	0.8	0.9	2.5	0.2	0	1.3	1.4	2.3	2
Sergey Karasev	22	12.3	1.3	3	0.433	0.7	1.8	0.4	0.6	1.2	0.481	0.552	0.4	0.4	1	0.6	0.9	1.5	0.8	0.4	0.1	0.5	1.9	3.8	2
Amedeo Della Valle	21	12	1.7	3.4	0.5	0.7	1.4	0.5	1	2	0.5	0.604	1.7	2	0.837	0.3	0.6	0.9	0.5	0.5	0	0.2	1.3	5.9	2
Paulius Jankunas	27	13.7	1	2.7	0.384	0.1	0.8	0.143	0.9	1.9	0.481	0.404	0.8	1.1	0.759	0.6	2.4	3.1	0.8	0.4	0.1	0.5	1.7	3	2

Cluster 3:

Player	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Daniel Hackett	23	20	2.8	6.7	0.418	1.5	3.3	0.442	1.3	3.3	0.395	0.529	1.3	1.6	0.806	0.7	1.9	2.6	2.7	1	0	1.5	2.8	8.3	3
Tyrese Rice	28	17.4	3	7.4	0.408	1.5	3.9	0.38	1.5	3.5	0.439	0.507	2.5	2.7	0.933	0.3	0.9	1.1	2.8	0.4	0	1.8	1.8	10	3
Victor Claver	17	21.2	2.1	4.1	0.507	0.9	1.9	0.455	1.2	2.1	0.556	0.616	0.7	0.7	1	0.9	2.1	3	1.6	0.8	0.2	0.4	1.9	5.7	3
Angelo Caloiaro	18	20.9	1.7	3.8	0.456	0.6	1.8	0.344	1.1	2	0.556	0.537	1.5	1.8	0.818	1	1.8	2.8	1.4	0.7	0	0.4	2.1	5.6	3
Johannes Voigtmann	28	22.3	2.9	6.1	0.474	1.5	3.6	0.41	1.4	2.5	0.563	0.594	0.4	0.8	0.571	1.2	4	5.2	1.5	0.6	0.4	1.1	2.4	7.7	3
Tyler Dorsey	28	18.9	3.3	8.6	0.388	1.9	4.7	0.394	1.5	3.9	0.38	0.496	1.5	2.3	0.683	0.5	1.9	2.4	1.5	1	0.1	1.5	2.1	10	3

Cluster 4:

Player	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Tarik Biberovic	6	10.7	1.2	4.2	0.28	0	1.2	0	1.2	3	0.389	0.28	0	0	0	0.8	0.8	1.7	1	0	0.2	0.5	1	2.3	4
Kalin Lucas	3	13	0.7	2.7	0.25	0	1	0	0.7	1.7	0.4	0.25	0	0	0	0	0.3	0.3	3	0.3	0	0.7	0.3	1.3	4
Vladislav Trushkin	24	11.1	1.3	3.6	0.349	0.9	2.8	0.328	0.3	0.8	0.421	0.477	0	0	0	0.6	0.7	1.3	0.3	0.1	0	0.2	1	3.4	4
Sergei Monia	17	11.3	0.6	1.7	0.379	0.6	1.4	0.435	0.1	0.4	0.167	0.552	0	0	0	0.5	1.4	1.9	0.5	0.3	0.2	0.5	1.8	1.9	4
Andy Rautins	9	7.7	0.8	1.6	0.5	0.7	1.3	0.5	0.1	0.2	0.5	0.714	0	0	0	0.2	0.1	0.3	0.2	0.3	0	0.1	1.9	2.2	4
Dogus Balbay	25	7.9	0.5	1.1	0.481	0.2	0.5	0.385	0.3	0.6	0.571	0.574	0	0	0	0.2	0.7	1	0.9	0.5	0	0.2	0.9	1.2	4

Cluster 5:

Player	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Michael Eric	28	15	2.5	4.5	0.552	0	0	0	2.5	4.5	0.552	0.552	0.8	1.6	0.457	1.2	1.5	2.7	0.3	0.8	0.6	1.3	2.5	5.7	5
Livio Jean Charles	27	21.1	3.6	7.2	0.505	0.6	1.8	0.313	3.1	5.4	0.568	0.544	1.3	1.4	0.872	1.6	2.2	3.7	0.6	0.4	0.7	1.1	1.9	9.1	5
Georgios Papagiannis	27	17.6	3.1	5.1	0.62	0	0.1	0	3.1	5	0.63	0.62	0.6	0.9	0.667	1.6	3.2	4.7	0.2	0.4	0.9	1	2	6.9	5
Gustavo Ayon	26	21.4	5.6	9.4	0.596	0	0	0	5.6	9.4	0.598	0.596	1.2	2.1	0.582	2.5	2.9	5.3	2.2	1.3	0.5	2.4	2.7	12.5	5
Greg Monroe	28	24.2	5.2	9.9	0.523	0	0	0	5.2	9.9	0.523	0.523	2.5	3.4	0.747	2.2	4.5	6.8	2.5	1.3	0.6	2.5	2.1	12.9	5
Zach LeDay	28	22.4	4	7.9	0.505	0.5	1.1	0.433	3.5	6.9	0.516	0.534	3.3	3.9	0.844	1.7	3	4.7	1	0.5	0.7	1.7	2.1	11.8	5

Cluster 6:

Player	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	CLUSTER
Peyton Siva	17	21.2	2.5	7.4	0.333	1.2	4.3	0.274	1.3	3.1	0.415	0.413	3.2	3.7	0.857	0.2	1.2	1.5	5.2	0.9	0	2.3	2.5	9.3	6
Shane Larkin	25	30	6.7	12.7	0.53	3.5	6.9	0.509	3.2	5.8	0.556	0.669	5.2	5.8	0.903	0.5	2.6	3.1	4.1	1.3	0	2.2	1.3	22.2	6
Mike James	28	28.6	6.5	14.9	0.44	2.8	6.7	0.42	3.7	8.1	0.456	0.535	5.2	6.2	0.833	0.5	2.9	3.3	4.3	0.7	0.1	3	1.6	21.1	6
Vassilis Spanoulis	22	24.6	3.8	9.3	0.41	1.5	5.4	0.286	2.3	3.9	0.581	0.493	2.1	2.6	0.807	0	1.2	1.2	4.6	0.4	0	2.6	1.5	11.3	6
Martin Hermannsson	27	25.8	4.1	10	0.41	1.1	3.2	0.337	3	6.9	0.443	0.463	1.6	1.9	0.86	0.3	1.3	1.6	4.8	0.5	0	2.7	2.1	10.9	6
Vasilije Micic	24	30.6	5	10.8	0.469	2.2	5.4	0.4	2.9	5.3	0.539	0.57	2.2	2.3	0.964	0.4	2.1	2.5	5.8	1.3	0	3.1	2.5	14.5	6