IS5126 HANDS-ON WITH BUSINESS ANALYTICS

Guided Project Report

Supervisor: DR TUAN Q. PHAN

Submitted by

Cho Zin Tun (Student ID:A0098996W, Email:e0230036@u.nus.edu)

Chua Hian Choon (Student ID:A0176643X, Email:e0232247@u.nus.edu)

Ngoh Chang Chiat Vincent (Student ID:A0176587J, Email:e0232191@u.nus.edu)

Toh Pei Xuan (Student ID:A0000584R, Email:e0229629@u.nus.edu)

Ye Honghai (Student ID:A0176590W, Email:e0232194@u.nus.edu)

1. Introduction

This report aims to establish a causal relationship on the factors that determine a player's high salary from the professional basketball data from http://www.basketball-reference.com. This information will be useful to basketball team owners in determining the salary amount that they should be paying new players based on the identified factors. The team has worked through the data pipeline focusing on the 3 focus area, namely web-scraping, performing ETL into SQLite and analyzing the data.

2. Web-scraping (Part 1)

Python library 'BeautifulSoup' and 'RegEx' were used to get the information about NBA players from https://www.basketball-reference.com/. Crawled Information includes players' stats-per tabs by season (Player Per Game, Player Totals, Player Per 36 Minutes, Player Per 100 Poss, Player Per Advanced), player's profile information, player's salary, basic team information, and team statistic by season.

To scrap for players' game statistic and salary by the game season, "https://www.basketball-reference.com/leagues" was being used as the starting url. After that, depending on the year and stats-per-tabs, corresponding url was being appended to crawl the respective data. For retrieving the players' and teams' information, "https://www.basketball-reference.com/" was used as the starting url with appended unique url for each player and team.

Below are the snippet of RegEx and BeautifulSoup codes for getting information from stats-per-tabs tables. Full version of the codes can be found in the submission folder.

```
### html is the object request from the statistic URL ###
tablesearch = 'id="'+ tableid+'"(.*\n.*)<\/table>'
table = re.findall(tablesearch,html,re.M|re.I|re.S)[0]
body = re.findall('<tbody>(.*\n.*)<\/tbody>',table,re.M|re.I|re.S)
trs = re.findall('<tr.*?>(.*?)<\/tr>',body[0],re.M|re.I|re.S)
for tr in trs:
    tds = re.findall('<td.*?>(.*?)<\/td>',tr,re.M|re.I|re.S)
    for td in tds:
        if 'players' in td:
            playerlink = 'https://www.basketball-reference.com' + re.findall('<a href="(.*?)">', td)[0]
        elif 'teams' in td:
            subteamlink = re.findall('<a href="(.*?)">',td)[0]
            teamlink = 'https://www.basketball-reference.com' + subteamlink
```

RegEx

```
### req is the object request from the statistic URL ###
text = BeautifulSoup(req.text, "html.parser")
stats = text.find('table', {'id': tableid})
for i in stats.tbody.find_all('tr'):
    for j in i.find_all('td'):
        text = j.get_text()
        if 'player' in str(j):
            playerURL = 'https://www.basketball-reference.com' + j.find_all('a')[0]['href']
        if 'team_id' in str(j):
            teamURL = 'https://www.basketball-reference.com' + j.find_all('a')[0]['href']
```

BeautifulSoup

When scraping for data within a table with properly formed html, BeautifulSoup is the more preferred way. Comparing to RegEx, BeautifulSoup code is easier to formulate with higher readability. On the other hand, RegEx requires deeper understanding of HTML tags used on the web site before the regular expression can be formulated but it gives the code more flexibility.

BeautifulSoup is used to scrap all the players' game statistics information from table since the html codes are properly formatted with attributes. For player profile and team basic information, RegEx is preferred. On player and team site, the information to be crawled for are placed behind html code without attributes or ID, using RegEx provides the flexibility that BeautifulSoup is lacking in, allowing more direct drilling down to the information needed.

Beside game statistic found in the season page, more in-depth breakdown on turnovers statistic were scrapped for the every players. This information was selected over the other in-depth

information (such as fouls breakdown or ball possession), because how a player lost the ball can be associated with how well the player handles ball and his teamwork with the team. Such information will be useful for team owner to build an analytic model.

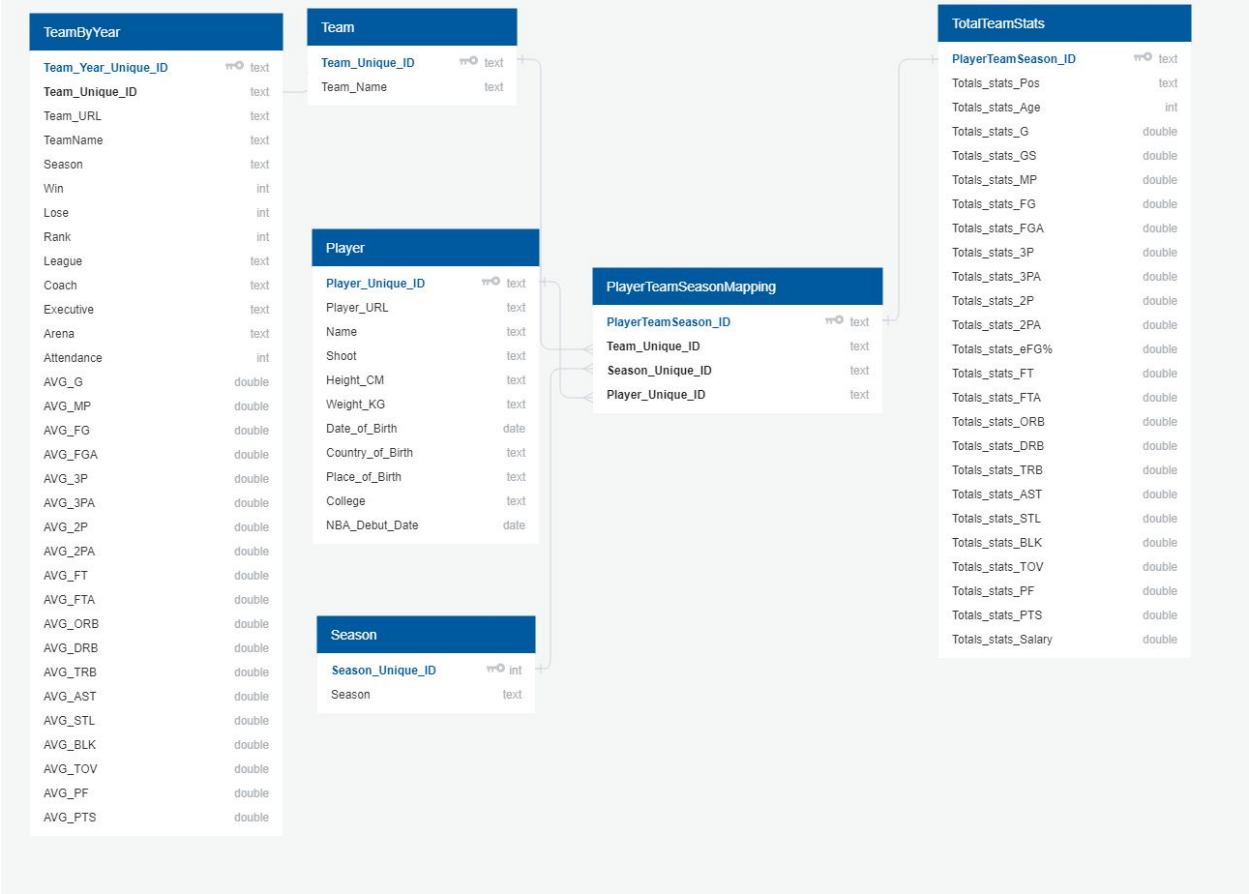
Another additional data crawled was the All-Star Roster table found on the season page. During each season, the cream of the crop are selected to represent their division (East or West) in a All-Star Game. These players are highly valued players of the season which team owners should keep an eye on and be ready to fork out high amount of salary if they want to recruit these players.

3. ETL (Part 2)

Normalization is performed to have tables in such a way that the table includes single valued attributes/columns, all the stored values in a column are of the same domain,all the columns in the table have unique names , and the order in which data is stored, does not matter. Partial dependency and transitive dependency were avoided as well during the schema creation.

Team, Player and Season tables were created separately and in each table, only those associated fields were hosted. PlayerTeamScoreMapping table was constructed to serve as a facade table for statistic tables. With this approach, whenever there is a modification (either deletion or update or insertion) for team or season or player, only one respective table is needed to be amended.

Below is the schema for the database:



Using the tables created on SQLite, various information can be obtained using different aggregation method in query string.

Information on active player of latest season (2016-17)

Item	**Stats**
Active Player Count	486
Average Age	26.4
Average Weight	99.3 kg
Average Experience	5.7 years
Average Salary in 2016-17	$6,170,918
Average Career Salary of Active player	$29,055,409

* Average Salary is based on unique player combination salary. Player who played in different teams will have their salary combined.

Information on Position Played of latest season (2016-17)

Position Played	Player Count
Center	98
Power Forward	103
Point Guard	97
Small Forward	84
Shooting Guard	105

* Position played is based on unique player's position. Player who played the same position in different teams is counted as 1 player.

Top 10%, bottom 10% and middle 50% based on players' salary in 2016-17

Only first 10 records of each query are shown. Full result can be found in result table submitted.

Top 10%				**Bottom 10%**				**Mid 50%**		
Player	**Team**	**Salary($)**		**Player**	**Team**	**Salary($)**		**Player**	**Team**	**Salary($)**
DeMarcus Cousins	SAC	33,915,800		Elijah Millsap	PHO	23,069		Kevon Looney	GSW	1,182,840
DeMarcus Cousins	NOP	33,915,800		Dahntay Jones	CLE	24,022		Skal Labissiere	SAC	1,188,840
LeBron James	CLE	30,963,450		Ben Bentil	DAL	31,969		Josh Huestis	OKC	1,191,480
Al Horford	BOS	26,540,100		Marcus Georges-Hunt	ORL	31,969		Kyle Anderson	SAS	1,192,080
DeMar DeRozan	TOR	26,540,100		Patricio Garino	ORL	31,969		Pascal Siakam	TOR	1,196,040
James Harden	HOU	26,540,100		Gary Payton	MIL	35,166		Larry Nance	LAL	1,207,680
Kevin Durant	GSW	26,540,100		Alex Poythress	PHI	38,903		C.J. Wilcox	ORL	1,209,680
Mike Conley	MEM	26,540,100		Justin Harper	PHI	57,672		Jordan Mickey	BOS	1,223,653
Russell Westbrook	OKC	26,540,100		Jarrod Uthoff	DAL	63,938		Luke Babbitt	MIA	1,227,000
Ersan Ilyasova	OKC	25,200,000		Axel Toupane	MIL	72,029		Brice Johnson	LAC	1,273,920

Information of current active players in previous seasons is as follows:

Season	Total Salary Paid ($)	Active Player in previous season	Average Salary ($)
2007 - 2008	440,268,722	95	4,634,407
2008 -2009	631,858,228	122	5,179,165
2009 -2010	750,902,479	151	4,972,864
2010 -2011	1,072,256,908	180	5,956,982
2011 -2012	1,089,045,835	207	5,261,090
2012 -2013	1,377,667,730	243	5,669,414
2013 -2014	1,637,961,962	284	5,767,471
2014 -2015	1,970,253,152	331	5,952,426
2015 -2016	2,151,647,313	386	5,574,215
2016 -2017	2,999,066,274	486	6,170,918

The average and variance of players' salary and years of experience of each team by season can be found in the submitted result tables. The cross tabulation format has been provided as well.

4. Analysis (Part 3)

Analysis was performed using player's performance data aggregated based on per game (like number of points scored and number of fouls committed) and players' details (like weight and whether they went to college). Please refer to full size images in folder 6 (analysis results) for images used in this section. Please also refer to the table in the appendix for glossary of the variable and short-forms used.

1) *K-Means*

K-Means clustering was done using data aggregated based on players. Variable selection on the data was done to remove factors and correlated variables based on domain knowledge (for example FG%=FG/FGA, where only FG (Field Goal) and FG% (Field Goal %) was included in the model, but not FGA (Field Goal Attempt). The data was also normalised before modelling to give all variables equal weights in the model. As *KMeans* method is based on distance, variables that are larger in magnitude, like Salary, will carry more weight than other variables, like FG%, if normalisation was not done.

Model was fitted based on multiple centers, where *K* is between 1 to 10. Each *K* was fitted 30 times (set option in *Kmeans* function of R as *nstart=30*) and the best model was chosen and reported. The best model is based on minimised total within Sum of Squares (WSS). Total WSS measures the compactness of the clusters and should be minimised to signify that data in the same cluster are similar.

Based on the fitted clusters, 3 clusters (*K*=3) was selected using the "elbow" method, where the reduction in total WSS tapers off after *K*=3 ("kink" in the plot represents minimal improvement in the model after *K*=3, see **figure 1** below for plot of total WSS against *K*).

Based on the model, the players can be grouped according to 4 variables, namely, number of games played, minutes played, points scored per game and salary of the player. The means of these variables in each cluster were the most variable and could represent the difference between each cluster (see **figure 2** below for means of each variable in the model based on the 3 clusters). The 3 clusters can be represented as "High", "Middle" and "Low" based on the 4 variables. In other words, players in the "High" cluster played in higher number of games, played for a longer time, scored higher points per game and have higher salary. This similarly applies to "Middle" and "Low" clusters where players are in the middle and low range of the 4 variables

respectively. Hence, fitter and higher performing players tend to command a higher salary (see **figure 3** for 3D plots of the relationship between the 3 clusters for MP, PS.G and Salary, other variable plots can be found in folder 6).

It is possible to obtain different clustering results when different starting centers are used to initialise the model. This can be shown in R by omitting the "nstart" option explained above and allow the model to run at one start point. The model is then ran for 100 times in total (each using a random start point) at each *K* (1 to 5). By ordering and plotting the centre of each cluster for each *K*, it is possible to show that this model is stable at *K*=1 to 3 as every centre (of the same 4 variable, number of games played, minutes played, points scored per game and salary of the player) plotted, falls onto the same point, which means that all 100 runs at random starting point is able to produce the same resulting 1 to 3 clusters (see **figure 4** where only 3 points can be seen on the graph for *K*=3, refer to folder 6 for graphs for *K*=1 and 2). Nevertheless, the model becomes unstable from *K*=4 and 5 at the same 4 variables (see **figure 5** and **6** below where multiple centre points can be seen for *K*=4 and 5) where all 100 runs are unable to reproduce the same set of results.

Based on the high, middle and low clustering, it is possible to deduce that players that played more and scored more are paid higher salary. The centers refers to the average of the variable of each cluster. For example, the highest cluster played 29.5 minutes per game on average and scored an average of 14.2 points per game. Nevertheless, these variables may be somewhat correlated as the longer player plays on the court, the more chance he has to score points. More analysis needs to be done to establish any casual relationship and the boss should be told that it may not be true that offering players higher salary will cause them to score more points to help the team win.

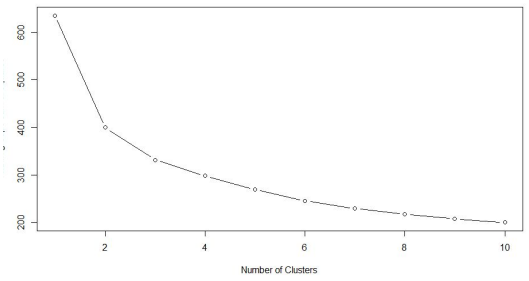


Figure 1: Graph of total within sum of squares for different *K*

	G	MP	FG	FG.	X3P	X3P.
	20.60662	9.052131	1.023758	0.3874699	0.2111856	0.1736148
	51.85906	19.036089	2.528352	0.4394776	0.5856425	0.2574016
	61.84861	29.542052	5.254652	0.4602685	0.8866847	0.2819446

	X2P	X2P.	eFG.	FT	FT.	DRB
	0.8131768	0.4152312	0.4202961	0.4421156	0.5511058	1.046555
	1.9434001	0.4718675	0.4861381	1.0444716	0.7150772	2.315322
	4.3699846	0.4864851	0.4978543	2.7862843	0.7625069	4.143180

	TRB	AST	STL	BLK	TOV	PF
	1.492568	0.6330947	0.2680428	0.1844457	0.5143756	0.980924
	3.143953	1.4614688	0.5773515	0.3503576	0.9728317	1.784243
	5.499876	3.0904602	0.9664194	0.6520594	1.9271292	2.301200

	PS.G	Salary	Height_CM	Weight_KG
	2.693811	1115403	200.5184	98.86985
	6.686065	2925196	200.3692	98.70110
	14.179248	8142735	201.5026	101.42932

Figure 2: Means of each variable based on the 3 clusters ->"G" - games played; "MP" - Minutes Played; "Salary" - Salary of player; "PS.G" - points scored per game

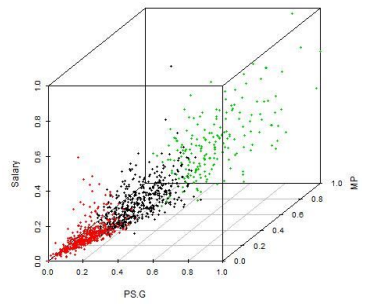


Figure 3: 3D plots based on MP, PS.G and Salary. 3 colour represents the different clusters.
Green - High; Black - Middle; Red - Low

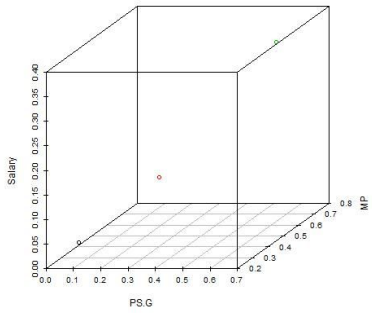


Figure 4: 3D plot of 100 centres for K=3 at MP, PS.G and Salary. There are only 3 points which means that all 100 runs (with random starting points) generated the same clustering result, and hence is stable

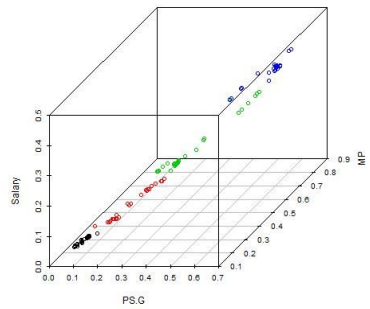


Figure 5: 3D plot of 100 centres for k=4. Multiple center points for 100 runs of kmeans (with random starting points).

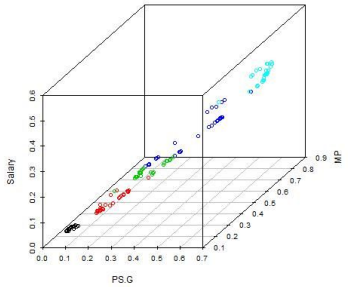


Figure 6: 3D plot of 100 centres for k=5. Multiple center points for 100 runs of kmeans (with random starting points).

2) *Linear Regression*

Linear Regression was run to examine which predictors affect the player's salary based on data from the current season only. Feature selection was performed by omitting variables, which were very sparse, such as Player Name, Date Of Birth, and Place of Birth. Variables that are vital to predict the salary were also chosen based on domain knowledge. Furthermore, correlation was checked to ensure that there was no multicollinearity between variables (drop correlated variables).

Variables input to fit the model is as follows:-

Salary~G + X3P + X3P. + X2P. + eFG. + FT + FT. + TRB + AST + STL + BLK + TOV + PF + PS.G + Height_CM + Weight_KG + Pos + Shoot + College

Per game data was chosen over other data as such data is able to give a general sense of a player's performance over other types of aggregation like per minute. There were also other variables that were not used for this model like advance performance data as it had many similar fields as the performance data from per game (correlated), hence only the performance data from per game was used.

Further selection of variables was done based on StepAIC that chose the best model based on goodness of fit criteria. This is done based on dropping variables one at a time and choosing the model with the best goodness of fit criteria. This continues until the original model is better than the model with dropped variable. Results from this model can be found in **figure 7** below. Significant variables are Age, GS, TRB, PS.G and Height_CM. These significant predictors are from both player profile information and performance statistic. The higher salary the player earned meant that the older he was (more experienced), the taller he was, the more games he started, the more rebound he caught and the more points he scored.

The result makes sense since the player can be regarded as good player if he scores more points to help the team win or can catch more rebounds to increase the chances of ball possession. Moreover, the reason why he had started in many games could be because he is a good player. Age is also a positively correlated variable against salary and makes sense because with experience, the player becomes better and can get higher pay. Also, one other mildly significant variable, personal foul, is negatively correlated with salary which is as expected as fouls negatively impact a team's chances of winning.

Using the result from the analysis model, a team owner can motivate his team players by emphasizing on the the importance of teamwork. Every assist to score points and repossession of ball through rebounds per game increases average salary by coefficients given in the results of the model, which is together more than the increase in solely scoring for the team. At the same time, the team owner can also deter players from making fouls in game with the model results, by explaining that players with more fouls tend to have lower salaries. By reducing fouls and increasing ball possession and points scored, the team are more likely to win which ultimately allows them a higher chance to win any monetary prizes (revenue) associated with winning any championships.

Even though the model's result in terms of the relationship between variables makes sense, we cannot solely believe on it. There could be other unobserved or omitted variables that could affect the salary as well. The variables used in the model may also not be independent against each other as the performance statistics are all based on related games and may have some underlying relationship, hence underlying assumptions for linear regression model may not hold. Other causal analysis methods could also be performed to establish and casual relationship between salary and other variables. Also, magnitude of the coefficients modelled seemed to be

very large, it does not seem to make sense that one more rebound caught per game would increase the player's salary by $930,024.

```
Call:
lm(formula = Salary ~ Age + GS + X3PA + X2P. + ORB + TRB + AST +
    BLK + PF + PS.G + Height_CM, data = data_lm)

Residuals:
      Min        1Q    Median        3Q       Max
-13475022  -2499311   -309994   2156276  15032640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32435081    5853555  -5.541 4.77e-08 ***
Age            365769      42512   8.604  < 2e-16 ***
GS              45487       9851   4.617 4.90e-06 ***
X3PA          -277976     165924  -1.675  0.09447 .
X2P.         -2388524    1615671  -1.478  0.13992
ORB          -1765756     621589  -2.841  0.00468 **
TRB            930024     222518   4.180 3.42e-05 ***
AST            373230     167294   2.231  0.02610 *
BLK           -814940     536018  -1.520  0.12902
PF           -1166056     355056  -3.284  0.00109 **
PS.G           497406      70378   7.068 5.05e-12 ***
Height_CM      123010      28936   4.251 2.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4103000 on 524 degrees of freedom
Multiple R-squared:  0.5805,    Adjusted R-squared:  0.5717
F-statistic: 65.92 on 11 and 524 DF,  p-value: < 2.2e-16
```

Figure 7: Linear Regression model after variable selection using stepAIC

3) *Panel Data*

Variables are selected in the same manner as in linear regression. Both fixed effect model and random effect model was fitted to determine which model is more suitable.

Fixed Effects Model

The Fixed Model has been establish as follows by controlling for "Player" and "Year" that may be related to Salary.

```
modplm1=plm(Salary ~ Pos + Shoot + College + Age + G + GS + X3PA + X3P. + X2P. +
eFG. + FTA + FT. + ORB + TRB + AST + STL + BLK + TOV + PF + PS.G + Height_CM +
Weight_KG, data=dataplm3, index=c("Player","Year"),model="within")
```

From the analysis (results in **figure 8** below), it showed that there are 7 time varying variables that are significant based on the p-value. The 7 time varying variables are Age, ORB, TRB, AST, STL, PF and PS.G. Time-invariant predictors are removed in a fixed effect model.

```
Call:
plm(formula = formula, data = dataplm3, model = "within", index = c("Player",
    "Year"))

Unbalanced Panel: n=1107, T=1-10, N=4627

Residuals :
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-13147101  -1120977         0         0   1075739  14367555

Coefficients :
        Estimate Std. Error t-value  Pr(>|t|)
PosPF   282348.7   274074.1  1.0302 0.3029914
PosPG    30555.2   609753.1  0.0501 0.9600369
PosSF   135080.2   429483.0  0.3145 0.7531463
PosSG   401820.8   510109.1  0.7877 0.4309165
Age     374874.6    25366.6 14.7783 < 2.2e-16 ***
G        -8780.7     3141.7 -2.7949 0.0052197 **
GS        6657.4     3431.8  1.9399 0.0524708 .
X3PA    227993.7    85093.9  2.6793 0.0074118 **
X3P.    -28972.2   480012.7 -0.0604 0.9518746
X2P.   1359514.9  1144737.8  1.1876 0.2350633
eFG.  -3107554.6  1292819.1 -2.4037 0.0162816 *
FTA    -186274.8   112894.4 -1.6500 0.0990344 .
FT.     565689.0   388500.4  1.4561 0.1454593
ORB   -1052716.0   235031.6 -4.4790 7.738e-06 ***
TRB     503278.4    93873.4  5.3612 8.800e-08 ***
AST     472468.1    95580.7  4.9431 8.049e-07 ***
STL   -1109207.0   293906.2 -3.7740 0.0001633 ***
BLK     458124.8   268751.8  1.7046 0.0883505 .
TOV      12422.6   216300.5  0.0574 0.9542043
PF     -675110.4   154644.8 -4.3656 1.305e-05 ***
PS.G    332113.9    49383.5  6.7252 2.040e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    3.9527e+16
Residual Sum of Squares: 3.0274e+16
R-Squared:      0.2341
Adj. R-Squared: -0.012591
F-statistic: 50.9276 on 21 and 3499 DF, p-value: < 2.22e-16
```

Figure 8:
Results of Fixed effect model by controlling for player and year

Random Effects Model

The Random Effect Model has also establish as follows by controlling for “Player” and “Year”. From the analysis (results given in **figure 9** below), it showed that there are 6 time varying variables and 2 time invariant variables that are significant based on the p-value. The 6 time varying variables are Age, ORB, TRB, AST, PF and PS.G. The significant time invariant variables are whether the player went to college and his weight.

```
Coefficients :
               Estimate  Std. Error t-value  Pr(>|t|)
(Intercept) -18168336.9   3164713.2 -5.7409 1.002e-08 ***
PosPF          137666.3    185644.8  0.7416  0.458393
PosPG          141389.1    409694.3  0.3451  0.730028
PosSF           85711.4    252289.6  0.3397  0.734072
PosSG          364059.4    312829.9  1.1638  0.244581
ShootRight    -341712.9    220649.4 -1.5487  0.121530
College1      -826391.2    162938.2 -5.0718 4.095e-07 ***
Age            306471.8     13293.1 23.0550 < 2.2e-16 ***
G               -8308.4      2708.7 -3.0672  0.002173 **
GS               8287.7      3000.9  2.7617  0.005772 **
X3PA            28616.2     56098.3  0.5101  0.610000
X3P.          -122689.4    404981.5 -0.3030  0.761941
X2P.          1139317.4    951294.9  1.1976  0.231115
eFG.         -2563469.0   1050915.1 -2.4393  0.014754 *
FTA             64817.2     79650.6  0.8138  0.415820
FT.            -20330.2    299332.7 -0.0679  0.945853
ORB          -1108543.8    181108.2 -6.1209 1.008e-09 ***
TRB            625976.1     70411.5  8.8903 < 2.2e-16 ***
AST            459264.3     69113.9  6.6450 3.384e-11 ***
STL           -572913.6    213687.6 -2.6811  0.007365 **
BLK            474133.2    187512.9  2.5285  0.011487 *
TOV             31377.3    174026.7  0.1803  0.856924
PF            -802448.9    112282.3 -7.1467 1.029e-12 ***
PS.G           356188.9     33738.5 10.5573 < 2.2e-16 ***
Height_CM       44268.4     15628.1  2.8326  0.004637 **
Weight_KG       37090.5      9394.1  3.9483 7.989e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    6.9753e+16
Residual Sum of Squares: 4.0566e+16
R-Squared:      0.41856
Adj. R-Squared: 0.4154
F-statistic: 132.416 on 25 and 4601 DF, p-value: < 2.22e-16
```

Figure 9: Results of Random effect model by controlling for player and year

With reference to R-Square, Random Effect Model showed the value is almost higher than Fixed Effects Model by 2 folds. However, from the Hausman Test, the p-value was < 0.05, hence Fixed Effects Model to be used instead.

Based on fixed effect model, the amount of salary recommended to be paid to existing top and bottom 10% players can be found in folder 6 (topplmpred.csv and worstplmpred.csv). 10 players from each category is extracted and displayed in the table below.

Top Category	**Actual Salary**	**Predicted Salary ($)**	**Worst Category**	**Actual Salary**	**Predicted Salary ($)**
LeBron James	30,963,450	21,610,660	Elijah Millsap	23,069	787,189
Al Horford	26,540,100	14,074,935	Dahntay Jones	24,022	4,759,711
DeMar DeRozan	26,540,100	12,852,939	Marcus Georges-Hunt	31,969	31,969
James Harden	26,540,100	18,558,194	Patricio Garino	31,969	31,969

Kevin Durant	26,540,100	16,462,023	Gary Payton	35,166	35,166
Mike Conley	26,540,100	13,202,612	Axel Toupane	36,015	12,088
Russell Westbrook	26,540,100	19,871,909	Alex Poythress	38,903	38,903
Dirk Nowitzki	25,000,000	17,699,437	Quinn Cook	55,946	55,946
Carmelo Anthony	24,559,380	21,617,550	Wayne Selden	57,544	57,544
Damian Lillard	24,328,425	10,042,730	Justin Harper	57,672	1,977,336

Based on the model and table, top players seemed to be overpaid, while those with the lowest salary seem to be underpaid. The team owners can try to recruit undervalued players, who are those that have much lower actual salary than the predicted salary. These are the players with the potential to do well but costs a lot lesser to hire. They can also reduce their spending on star players or hire lesser of them to reduce their costs as they may not be worth as much as they are currently paid.

4) *Interpretation*

The above regression models could probably be suffering from endogeneity bias in the form of omitted variable bias. A positive relationship between omitted variable with both the predictors (*X*) and response (*Y*) variable will lead to the estimated coefficient of the predictors being larger than that of the true value. One example is experience or the number of years the player had played basketball. Experience affects the Salary (*Y*) and many performance indicator variables (*X*) positively and may be an omitted variable.

From the above established Fixed Effects Model, one could estimate treatment effect that are exogenous within each unit, we could hence perform causal analysis on the salary. As Fixed Effects Model on the panel data exploits within-group variation over time, we could perform causal relationship. By obtaining more observations about each player in each year and looking

at the effect of salary within each player in each year, the fixed effects regression could remove the effect of omitted variable bias.

Nevertheless, to truly believe that omitted variable bias have been removed from the fixed effect model, a key assumption is that there are no changes in Salary of the players that had not be control for. However, this assumption is not always true. Salary could be affected by injuries or market forces at the point of contract, that cannot be controlled for. The system of signing contracts in NBA would also likely affect salary. Contracts are signed for a period of time and will remain unchanged throughout that period. Hence performance or models modelled above may not always reflect the actual situation.

Causal analysis can be performed using Instrumental Variable (IV) analysis based on controlling for “experience” as mentioned above. “Experience” in terms of the number of years in playing basketball affects both salary and performance indicators used in the models above. By controlling for “experience”, the biases introduced by endogeneity could be avoided.

Another possible method to account for causal analysis of salary could be to split players into 2 groups, treated group is with players those signed contract during that year, while the control group consists of players who did not sign the contract that year. As it is unlikely that players can choose the year their contract is signed, regression discontinuity could be performed to exploit exogenous characteristics of the intervention to elicit causal effects. In this case, if there is a difference between the two groups of salary after controlling for the grouping, it could be concluded that the signing of contract year tend to have an effect on Salary.

All predictions above is based on a single player, but winning a game is a team effort. Hence, short-term selfish interest can be at the expense of collective objective of winning the game. This can be solved by including a high weightage of Assist and Assist Percentage in the prediction to maximise the objective of win. This may encourage players to help each other and

cooperate more to score points rather than always trying to score even though he may not be in the best position to do so.

If new athletes' contract are offered in accordance to teams in randomly set order, players might more likely accept contracts with teams at the tail end of the selection system as there would not be any more offers made if they were to reject the last few teams. Hence teams could consider paying lesser to recruit players at the end of the selection. As a result, those at the front end of the selection process might need to offer higher salaries to get good players to sign as compared to teams at the tail end of the selection process.

Based on current demographics of the team, the team may reject players. For example, if there are more than enough players playing the forward position, then the team may not be able to recruit another forward position player. Also, if the team focuses on being defensive, then only players with good defensive performance indicators may be a better fit for the team. The player could vice versa also reject a team's offer based on things other than salary. One example is if the player already had a team in mind where the player's role model is playing, then the player may only wish to accept offers from that team.

Appendix

Glossary
Rk -- Rank
Pos -- Position
Age -- Age of Player at the start of February 1st of that season.
Tm -- Team
G -- Games
GS -- Games Started
MP -- Minutes Played Per Game
FG -- Field Goals Per Game
FGA -- Field Goal Attempts Per Game
FG% (FG.) -- Field Goal Percentage
3P (X3P) -- 3-Point Field Goals Per Game
3PA (X3PA) -- 3-Point Field Goal Attempts Per Game
3P% (X3P.) -- FG% on 3-Pt FGAs.
2P (X2P) -- 2-Point Field Goals Per Game
2PA (X2PA)-- 2-Point Field Goal Attempts Per Game
2P% (X2P.) -- FG% on 2-Pt FGAs.
eFG% (eFG.) -- Effective Field Goal Percentage. This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.
FT -- Free Throws Per Game
FTA -- Free Throw Attempts Per Game
FT% (FT.) -- Free Throw Percentage
ORB -- Offensive Rebounds Per Game
DRB -- Defensive Rebounds Per Game

TRB -- Total Rebounds Per Game
AST -- Assists Per Game
STL -- Steals Per Game
BLK -- Blocks Per Game
TOV -- Turnovers Per Game
PF -- Personal Fouls Per Game
PS/G (PS.G) -- Points Per Game