IS - 5126 HANDS ON WITH BUSINESS ANALYTICS

GUIDED PROJECT - 2 Team 06

Jayakumar Alagappan Meenakshi	A0134431U
Krishna Kannan	A0134475A
Raghavendhra Balaraman	A0123443R
Vishnu Gowthem Thangaraj	A0134525L

Data Preparation :

The salary of all the players from basketball-reference.com is spread over a wide range. As a result, when these values are plotted , we find that salary curve is skewed toward the left. In order to remove this skewness, log(Salary) is calculated to fit the curve in the normal distribution as shown in Figure 1. Thus, we will use log(Sal) throughout our analysis.

Figure - 1

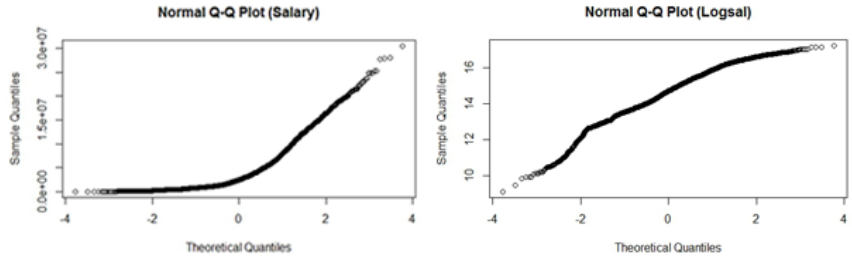
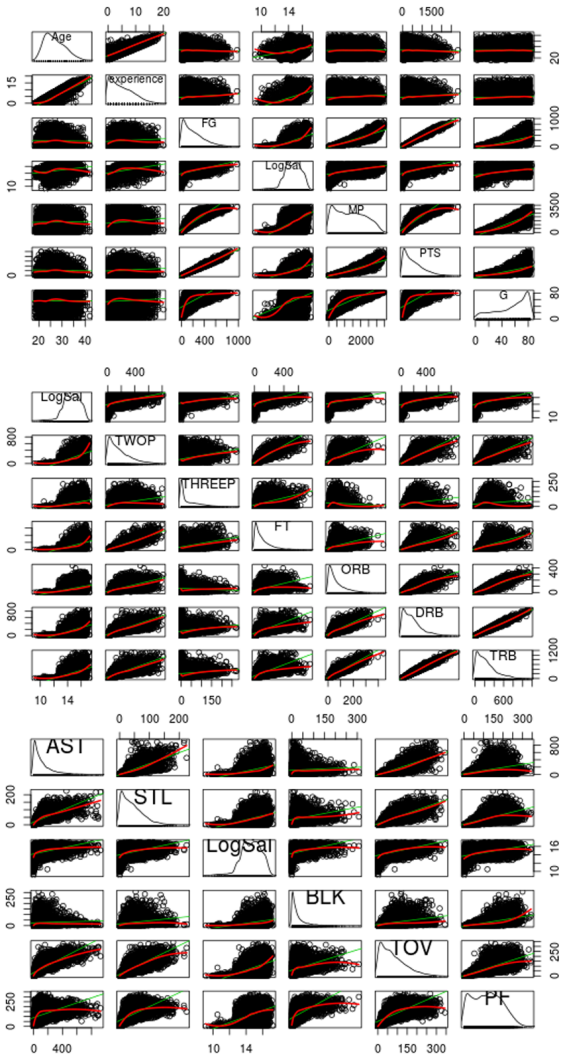


Figure - 2



Identifying Relevant Parameters :

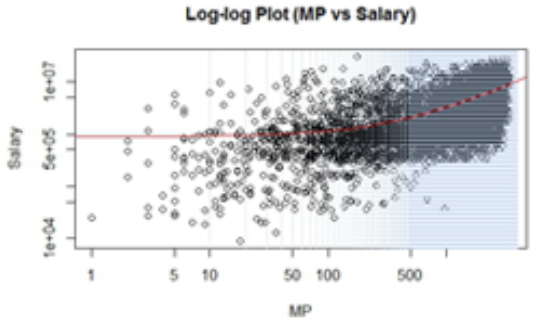
In this phase of analysis, we have identified those parameters that are more related to salary. This analysis is performed using the scatter plot functions in R. The below plots (Fig 2) are the scatterplots diagrams for different parameter plotted against salary.

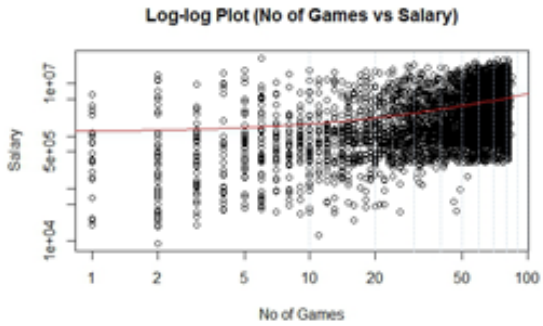
From the scatterplot diagram, we understand that MP (Minutes Played), PTS (Points), G(Total Games), 2P and 3P are very linear to the salary. This explain that these variables could be highly related to salary. In addition, parameters like experience, FG(field goals), FT(Field Throws) and TRB(Total Rebounds) also show significant relationship to salary. Further, the scatterplot diagram indicates that the salary is high for players in the age group of 25-30 yrs. As a result, we can observe that the salary of the players increases until 6 yrs of experience and gradually decreases after that.

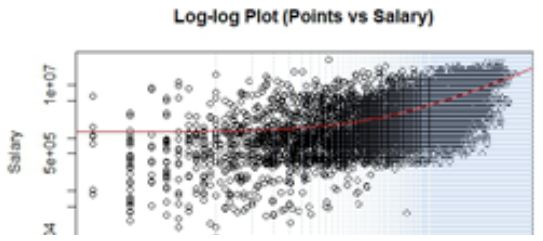
Q1) K Means Clustering :

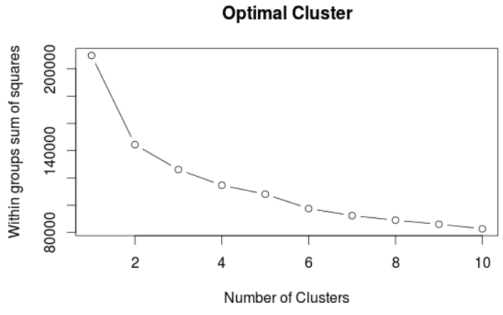
In performing the cluster analysis, the following parameters

are being considered to form the cluster. MP (Minutes Played) : This is one of main criteria as it tests the stamina and fitness of a player as the players are supposed to run between the courts for a longer duration. More the minutes played, higher is his stamina and he becomes a more sought out player. PTS (Points): Points plays a significant role in sports. Players are required to maintain a good average of points over the season to increase their salaries. G (Games) : Higher the no of games played, more experienced a player is. 2P (2 pointers) and 3P (3 pointers): These two add to the points scored by a player. They reflect a players potential to score points. TRB (Total Rebounds) : This is important parameter for any defensive players. The selection of these parameters are justified using the log-log plot plotted for each of the above parameter against the salary. We see these parameters are linear to the salary. We decided to cluster the dataset with the above described parameters for different centers. In order to identify the no of clusters required to split the dataset, 'elbow method' was used. The below graph plot show at k=3, the curve starts to decline gradually. So, we decided to partition into 3 clusters. In order to prove that 3 clusters are enough to partition the dataset, cluster summary with k= 2,3,4 was used to analyse.









From the below cluster summary, the betweenss/totss for k=3, 4 are 0.8445956, 0.8950121 respectively. The difference between k=3 and k=4 is very small and moreover the clusters formed when K=3 is by far more understandable than k=4, as 3 clusters nicely divides the salary into three main regions i.e., Low, Medium and High.

K=3	MP	PTS	G	TWOP	THREEP	Experience	TRB
1	2583.95	1213.90	76.2154	377.700	71.3248	5.513421	449.26084
2	439.988	137.24	34.573	44.3567	8.29603	4.351587	78.42738
3	1568.39	565.377	68.085	175.011	38.5689	5.057341	268.05141
Betweens s/totss	0.8445956						

K=4	MP	PTS	G	TWOP	THREEP	Experience	TRB
1	2767.56	1383.75	77.35	430.073	79.1402	5.559783	480.92283
2	315.298	95.6716	28.7934	30.8041	5.8001	4.137685	55.50994
3	1157.37	391.921	61.4471	124.01	25.3402	4.96873	202.46967
4	1992.75	769.166	72.3753	237.637	51.3614	5.299472	341.44327
Betweens s/totss	0.8950121						

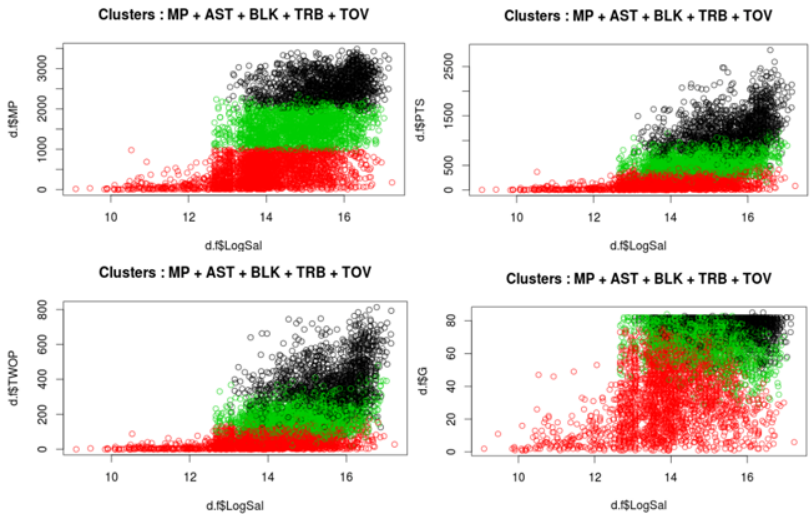
Interpretation to the boss :

The cluster formed with 3 centres considering different parameter is now plotted for each of the independent variables against the salary. We observed that, MP (Minutes Played), G (Games), PTS (Points), 2P partition the dataset into 3 clusters significantly. Based on these analysis following are some of the interpretations.

From Fig 1, we see that MP (Minutes Played) separates the clusters more accurately than any other parameters. This infers that players with low salary are usually the one who plays less than 1000 minutes in an average. Interestingly, there are many high paid players who still play for less than 1000 min. These players could be the senior most player in a team who has more experience but do not play in small tournaments like league or participate only in some crucial matches.

From Fig 2, we observe that players with low salary are the one who have scored less than 500 points on an average. Players with more than 1500 points lie in the high salary category. Here again we see some of the players with low points earn high. This could be because a franchise might have paid a hefty amount to a player with lots of expectations on him. Unfortunately it could be a bad season for that player or suffered with any injuries. Still the trend continue to show that players who is a high scorer is paid high.

From Fig 3, we see large number of highly paid players are the one who has scored more than 400 two pointers on an average. The large populated black areas also reveals that players often score 2 pointers in a match.



From Fig 4, we can observe that the players who have played around 50 games in an average are more decently paid. On the other hand, high paid players are the one who have played more than 80

games on an average. This clearly infers that total no of games played by any player seems more important in deciding his salary.

Q2) Linear Regression

The preliminary investigation conducted to identify the correlation between salary and the below mentioned quantitative independent variables show a comparatively high positive correlation towards salary. **We understand the salary of any player does not depend on current years performance, but rather on previous years' performance. That is, salary contracts are made before the start of the season and depend on previous years' performance with the preceding year's performance being most influential. Thus, the linear model fits the data ranging from seasons 2000 to 2011 as we are orchestrating to predict a player's salary in the season 2011-2012.** From the model, the **stars are shorthand for consequentiality levels**, with the quantity of asterisks shown according to the p-worth figured. In this case, the asterisks ('***', '*') depict that there is a likely **significant relationship** between the player salary and the parameters - **Experience, Points, Total Rebounds(TRB), Turnovers(TOV), Block(BLK).** The **residuals** are the contrast between the authentic values of the variable we are foreseeing (i.e. Player Salary) and the anticipated values from our regression. For most regressions the residuals resemble a normal distribution. The residual mean values closer to 0 demonstrate that the model exhibits a superior fit towards the actual data points. This model is discovered to be measurably critical, with **Adjusted R-Squared** (It is the proportion of the variance in the data that's expounded by the model.) accomplishing very much an adequate estimation of around **52%**. (Higher score is better with 1 being the best.). The **Coefficients - Estimate** is the value of slope computed by the regression. This number will conspicuously vary and depends on the magnitude of

```
Call:
lm(formula = LogSal ~ experience + FG + PTS + FT + TRB + PF +
    TOV + BLK, data = data.f)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0852 -0.4770  0.0361  0.5509  2.4602

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.1737017  0.0265629 495.944  < 2e-16 ***
experience   0.1201670  0.0030425  39.497  < 2e-16 ***
FG          -0.0019388  0.0008784  -2.207  0.02735 *
PTS          0.0015725  0.0003713   4.235 2.33e-05 ***
FT          -0.0010121  0.0004080  -2.481  0.01315 *
TRB          0.0006956  0.0001486   4.681 2.94e-06 ***
PF           0.0008380  0.0003185   2.631  0.00854 **
TOV          0.0021895  0.0004837   4.526 6.14e-06 ***
BLK          0.0022404  0.0005264   4.256 2.12e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8249 on 4725 degrees of freedom
Multiple R-squared:  0.5179, Adjusted R-squared:  0.5171
F-statistic: 634.4 on 8 and 4725 DF,  p-value: < 2.2e-16
```

Dependent	Independent	Correlation Value
Salary	Experience	0.458
	Field Goals	0.569
	Minutes Played	0.574
	Points	0.566
	Games	0.446
	2P	0.553
	3P	0.307
	Free Throws	0.514
	Total Rebounds	0.516
	Personal Fouls	0.486
	Turnovers	0.539

the variable inputted into the regression.

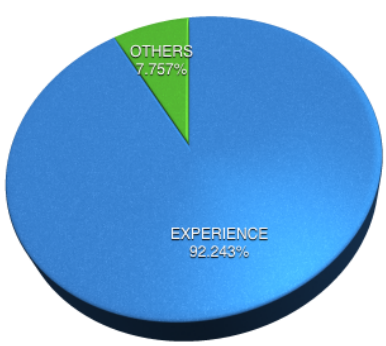
Interpreting the Model :

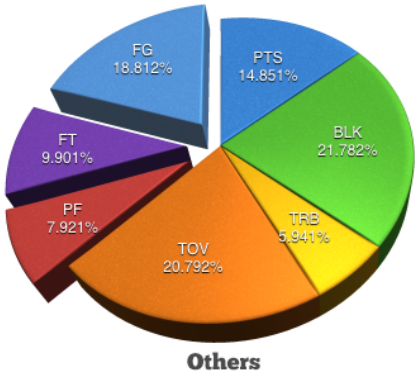
Regression Equation

log(salary) = 13.17 + (0.1201 * experience) - (0.0019 * FG) + (0.0015 * PTS) - (0.0010 * FT) + (0.00069 * TRB) - (0.00083*PF) + (0.00218*TOV) + (0.0021*BLK).

From the regression model, we observe that the magnitude of every variable contributes to average salary of an player. This can be interpreted as follows

For every unit increase in the experience of a player, the average salary increases by a magnitude of 0.1201 units. Similarly, this relation can be extended to other parameters too. As a result, for every increase in PTS, TRB, TOV, BLK , the magnitude of the salary increases by 0.0015, 0.00069, 0.00218, 0.0021 respectively. This interpretation can be further extended and plotted as pie chart to arrive at a definite conclusion. During the regression analysis, the average salary provided to any player in 2011-2012 season is calculated as $4,221,073. From the above plot, we observe that the regression model is deeply inclined towards the experience of a player, as it contributes a major factor in predicting a players salary around (92%). The rest of parameters including PTS, BLK, TRB, TOV, FG, FT, PF contribute to the remaining 8% of the salary. However, parameters FG, FT, PF (shown as shifted away from the circle), contribute negatively to the overall salary. In short, the following provides an in depth analysis about the model.





Salary per parameter :

The mean experience of the players is calculated to be around 5 yrs. In addition, the mean salary experience alone contributes to $3,893,644. **Therefore for every one yr increase in experience, the salary of a player increases by $780,289.** The total points scored by any player contributes to 14% (from the above chart) of Others Factor. During analysis, the average points scored is estimated around 421. **Thus, for every single point scored in a game by a player, the salary increases by $115.** Similarly, **for every rebounds, including Offensive and Defensive, the salary increase by %105.71**, From the regression model, we observe Blocks(BLK) are a significant predictor of salary which is confirmed from the above chart, as it contributes 21.7%

of the 'Other' factor. **On analysis, we have estimated that every block against an opponent, a player earns $3274.26 as his salary, Interestingly, for every Personal foul made by a player results in a decrease in salary by $305.693.** Even though a player is very skilled, a single foul could cost his salary.

Do you believe in results ?

From the results obtained in the regression model, we tend to believe that experience is the only major factor that plays a major role in deciding a players salary as it contributes 92% to the overall salary. However, the scatter-plots gave a different interpretation for the same. There, the salary of the player increases with increase in experience until 6yrs and later gradually decreases. Since this plot was not completely linear, one cannot conclude that experience alone plays a big part. In addition, in the beginning of analysis, the correlation against each parameter was calculated. The top 5 highly correlated parameters are MP (Minutes Played), Field Goals (FG), PTS(Points), 2P, TRB (Total Rebounds) and experience being one of the least correlated parameter. On the contrary, MP, FG and FT was calculated as the least significant parameters through this model.

Panel Data Analysis

Using Panel Data Analysis to predict the salaries of basketball players over the past ten years, data was collected from basketball-references.com for a ten-year period for all players during the time. Over the 10 seasons, a lot of changes may have happened in the careers of basketball players, potential metrics that could be used to predict their salary. For instance a player may have moved to another team or he may have been sidelined for a few years due to an injury and could have returned to play in subsequent seasons. There could also be playing strategy related decisions, where a player could have changed his playing position from one to another. Also, with increasing experience over the seasons, a player's performance could have correspondingly improved resulting in an increase in salary. Also, this data is a good fit for Panel Data Analysis as usually for players data(Individual data) is aggregated over seasons(Time). The overall metrics for basketball players across the seasons can be broadly classified into two categories.

Time Invariant Variables - Height, Weight, Playerid. Certain parameters with respect to a basketball player's career would always remain the same and never change across seasons.

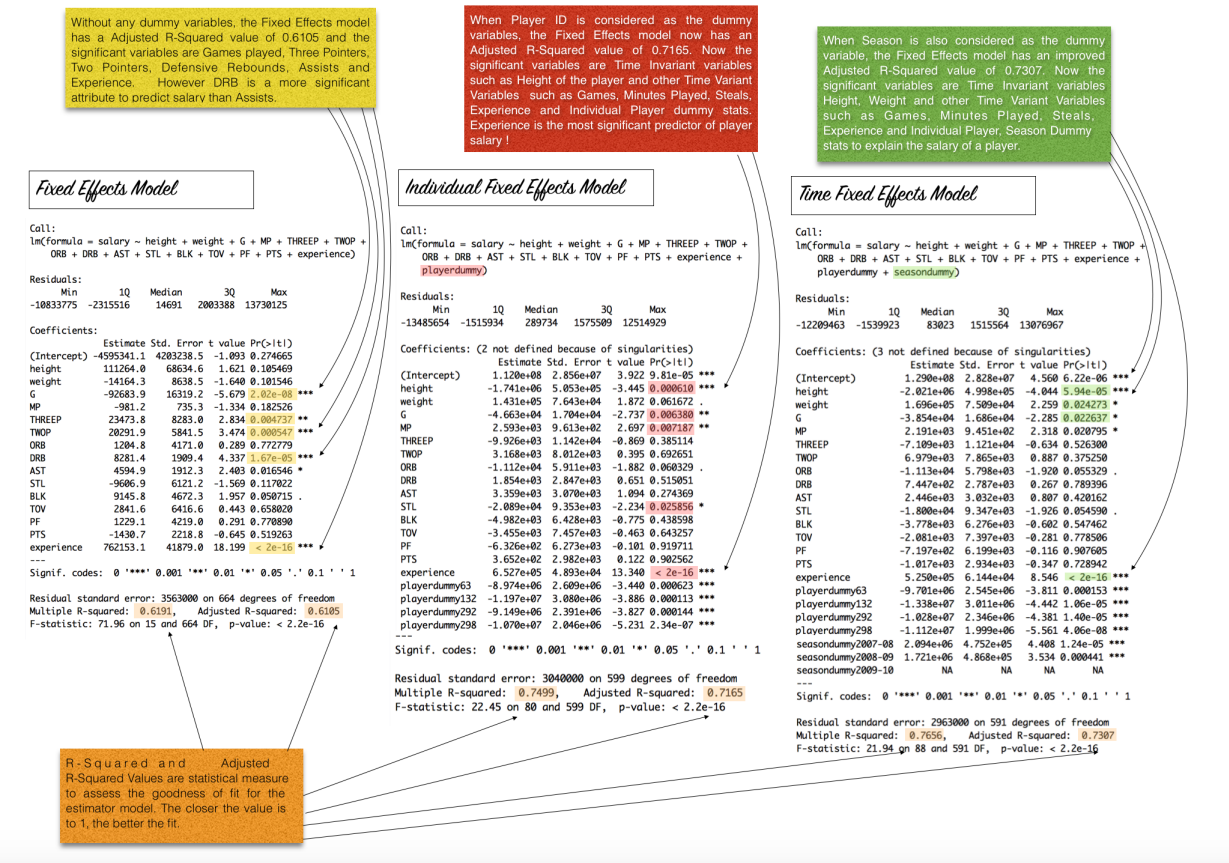
Time Variant Variables - G, MP, THREEP, TWOP, ORB, DRB, AST, STL, BLK, TOV, PF, PTS, experience. Every other variable that varies across seasons and could have an impact on the salary of a basketball player can actually be considered as a Time Variant Variable. The Total Rebounds (TRB) or Field Goals (FG) was not considered as they would exhibit multicollinearity as a result of their high correlation with ORB, DRB and TWOP, THREEP respectively.

Time Indicator - Season

We have considered a **balanced data for panel data analysis**, which means that only players who have played all the seasons during the 10 season period is considered for analysis. Same practices and theory could be slightly modified for the unbalanced data analysis as well.

INDIVIDUAL AND TIME FIXED EFFECTS MODEL (With/Without Dummy Variables) :

Over the balanced basketball players panel data, We have only considered Fixed and Random Effect models for our panel data analysis. Fixed Effects Model aims to explain the outcome variables (the salary in our case) in an improved fashion, under the assumption that each individual (the players in our case) has his own factors to affect salary. Only time invariant variables have this property. This leads to the inclusion of dummy variables in the data set, where Fixed Effects are in turn subdivided into Individual Fixed Effects and Time Fixed Effects models. When both are compared, the adjusted R square value was found to be more accurate for Time Fixed Effects model.



FIXED VS RANDOM EFFECTS MODEL :

Fixed Effects and Random Effects are very popular estimators with respect the panel data. By running that Hausman test, it was clear the Fixed Effect model was found to be consistent and the Random Effect model was found to be inconsistent. The high idiosyncratic value also means that the Random Effect is erroneous and

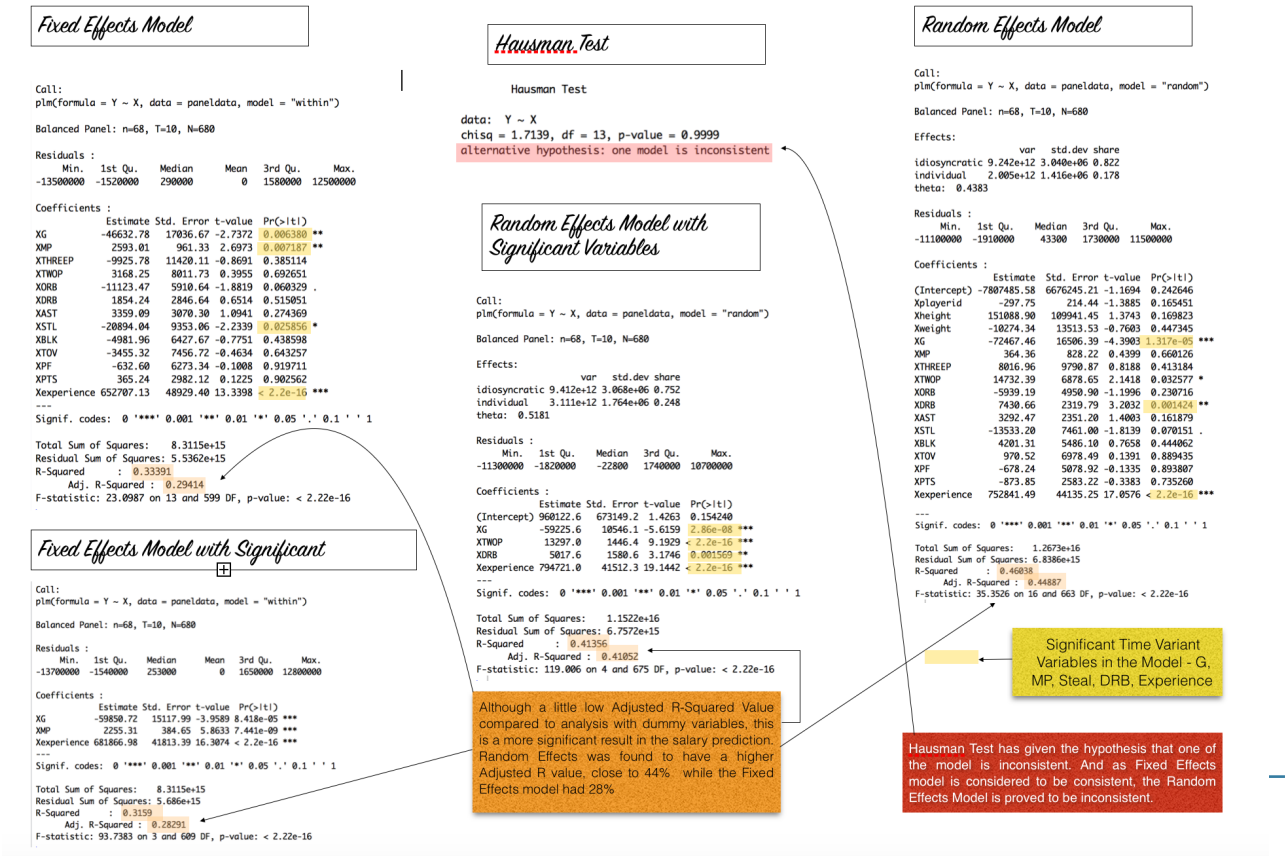
inconsistent. The Fixed vs Random chart consists of the Random vs Effects model and their corresponding analysis

SALARY PREDICTION AND INSIGHTS :

Across all models, Experience was found to be a significant predictor of salary. An experienced player playing in the NBA are more likely to be a top earner as well.

And although, it might come as a shock, height also had a negative effect on Salary. Basketball players, so often their skills attributed to their tallness or lankiness by the general public are actually not earning a lot of money because of their height. The reason is perhaps, almost all basketball players are very tall (Average basketball players height has been 6'6" to 6'8" for the past two decades) and so, if a basketball player is shorter, he must have been a very good basketball player to be in the team, thereby earning a lot more. Retired player, Allen Iverson for the Philadelphia 76ers is a perfect example of this.

Salary can be predicted using the model estimate and co-efficent values found from the analysis. Experience and the minutes played in the season for a player are highly valued in the Fixed effects model for salary prediction at 681866.98 and 2255.31 coefficient estimates, whereas the Games played had a negative impact coefficient estimate of 59850.72 units. Assists had a positive impact for random effects model with a coefficient estimate of 7430.66. We correspondingly advise the team owners (with the salary of current players based on the significant predictors discussed above. Steals have a negative indicator which could mean that, although being an asset to



have the ability to steal the ball from the opponent, it could not show on the salary the player takes home after the match. Perhaps players who steal the ball a lot more are underrated in NBA.

Q4 Interpretation

The confidence intervals were plotted for the Fixed Effects Model to predict player salary from Panel Data Analysis. All the significant salary predictors such as Experience, Minutes played or Games played does not have an upper bound or lower bound that consists of 0. Thus the slopes of the salary predictors are more significant contributing to different salary variations.

```
> confint(fixedestimator)
                    2.5 %      97.5 %
XG          -80024.0286 -13241.5215
XMP            708.8264   4477.1840
XTHREEP     -32308.7726  12457.2221
XTWOP       -12534.4494  18870.9402
XORB        -22708.1174    461.1719
XDRB         -3725.0629   7433.5512
XAST         -2658.5897   9376.7636
XSTL        -39225.6968  -2562.3899
XBLK        -17579.9674   7616.0423
XTOV        -18070.2269  11159.5778
XPF         -12928.1248  11662.9217
XPTS         -5479.6130   6210.1006
Xexperience 556807.2768 748606.9897
```

Data Selection: Scraped data of players/teams from 2000 to 2014. Select non-performance based parameters like experience, University/College.

Analysis: Root cause analysis techniques like Pareto Analysis, Fish Bone Diagram, and Histogram, and so forth can be utilized.

Causal Modeling of Experience vs Salary - In the event that you pay peanuts do you get monkeys ?

log(salary) = 13.17 + (0.1201 * experience) - (0.0019 * FG) + (0.0015 * PTS) - (0.0010 * FT) + (0.00069 * TRB) - (0.00083*PF) + (0.00218*TOV) + (0.0021*BLK). From the regression model (Refer section 2), for a year increase in experience, the salary of a player increases by 92%.

Why experience affects salary ?

As players age, it is conjectured that their pay rates increment, essentially as they gain more experience that can be considered independently from their recent performance. It is likewise estimated that as a player ages, their execution is superseded by their "body of work" and their income therefore depends more on their past performance rather than their latest statistical performance. From the scatter plot matrix, we observed that the salary of a player increases until an experience of 6 yrs after that the salary gradually declines. These senior most players could be regarded as mentors/coach to any teams. Such players are neither purchased at very high nor low price. Thus, they could be purchased at a very decent price for any low budgeted teams as a senior player in their team

The four main factors in basketball that helps in winnings games for teams include 1) Shooting 2) Turnover 3) Rebounding 4) Free Throws. Each of the above parameters are measured and tracked as a team instead of individual players. Most of the shooting attempts to the basket primarily comes with the help of the teammates. Thus teams with players having high 2P, 3P, TOV, TRB have a high chance of winning a game

Also, the defensive rebounds are taken from the teammates whereas offensive rebounds are solely depend on the players. Thus, teams can have efficient players to play the defense. This increases a greater chance of taking the offensive rebounds from the opponent team

A players accurate shooting into the basket largely depends on the number of assists from his teammates. Thus, a team should have a player who provides better assist so that his teammates shoot better and quite often.

Assigning appropriate players at appropriate positions would largely matter for a win. Example, Center and Point Forward (P-F) are the positions where the players get maximum rebounds and those players standing in that position should ensure that they do not commit turnovers. Naturally, every team should have big men in the Guards (G) position. However, unlike football or cricket, different analysis may assign positions differently for the same team

i) In order to buys a fresh player into a team, the following are some of the common ways in which the players are chosen by the team. There can be good affiliation between the team and the university in which the players study. The historical data would determine a large number of players in a franchise from a particular university. Such offers are mostly not rejected by those university players.

ii) The initial pay of a player can be determined using the player's past performance in the college or high school level games. These parameters are applied in the linear regression parameter against the budget available of the team during the season. The regression equations above are of much use here to arrive at an estimation. Thus, if the total budget of a team is around 10 million USD, then the salary of 'Ray Allen' during 2013-14 can be computed by calculating the average experience, FG, PTS, FT, TRB, ,PG , TOV and BLK of the player during his past years. These average metrics are then substituted against the above equation to arrive at a player's average salary. Let us assume the total salary for a player obtained from the above equation is 'X'. Thus a team owner can offer a initial salary 'X'.

iii) Besides overpaying/underpaying , some of the common reasons a team could reject players include,

- The player might not have popularity and do not have merchandise.
- The player may not be a crowd puller or in other words he might not be a known face among the general public inspite of his good performance records
- The player might be from a rival team or high school
- One common reason that was more prevalent during 70s and 80s was racial discrimination.