School of Computing

FACTORS AFFECTING GAME SALES

- Reshma Jawale

Table of Contents

Introduction 3

Part A: Web Scraping 3

Objective 3

Part B: Data Preparation and Description 6

Part C: Data Analytics 9

Clustering 9

Linear Regression 11

Model Justification 14

Panel Analysis 16

Interpretation 21

Appendix A 22

Data Dictionary for Raw Data 23

Appendix B 25

Appendix C 25

Introduction

Part A: Web Scraping

Objective

In this section we grab information of all NBA players who were active during the 2009/10 season to 2020/21 season. The task has been split up into 2 parts and performed separately.
Part 1 - scraping players statistics
Part 2 - scraping team statistics
We answer the questions below based on the different approaches used for each part.

1. Identify one or more starting URLs to crawl and explain your scraper's workflow.

Part 1 - Player statistics:
To go through the information of active NBA players from seasons 2009-2021, on the https://www.basketball-reference.com/ website we navigate to the Seasons header. Next, we selected any one season -> Player Stats -> Totals to scrape ALL players information for that season.
For scrapping, we have divided our url into two parts.

```
url_part1 = 'https://www.basketball-reference.com/leagues/NBA_'
url_part2 = '_totals.html'
```

We create a loop to go through each season year, and combine both urls as

```
url = url_part1 + str(years[i]) + url_part2
```

Part 2 - Team statistics:
We use a scrapy spider in order to get all the team statistics for the seasons 2009-2021. We use the following as the start URL: https://www.basketball-reference.com/teams/
This link has all the teams indexed. The spider scrapes the link for each team on the page then follows them one by one. On the team page the spider gathers all the basic information of the team and then finds the links to all the seasons played by the team. It then loops over each link specified in the time-period and gets the season statistics for the team.
Example flow of the spider:
Teams List page → Atlanta Hawks franchise page → Atlanta hawks 2020-21 season page

2. Identify the links to follow using either BeautifulSoup 4, or CSS/XPath and provide your code to do so.

Part 1 - Player statistics:

After we get the URL of the page, we create a variable ‘r’ and set it equal to

```
r = requests.get(url).text
```

Now we have a response object called r. We can get all the information we need from this object. Requests will automatically decode content from the server. Most unicode charsets are seamlessly decoded. When we make a request, Requests makes educated guesses about the encoding of the response based on the HTTP headers. The text encoding guessed by Requests is used when we access variable soup mentioned below.

The variable Soup holds the parsed HTML page of the URL we are extracting data from.

```
soup = BeautifulSoup(r, 'lxml')
```

Part 2 - Team statistics:

We start from the teams index page. Here we can scrape the team name and its link using the following code:

```
teams_list = response.xpath('//*[@id="teams_active"]/tbody/tr/th/a').getall()
```

Output gives a list containing team links as follows: `['<a href="/teams/ATL/">Atlanta Hawks</a>']`. Which can be split using regex to get the team name and link

From each team link we get the season link from the following code:

```
links = response.xpath('//*[@id="' + team_id + '"]/tbody/tr/th/a/@href').getall()
seasons = response.xpath('//*[@id="' + team_id +'"]/tbody/tr/th/a/text()').getall()
```

Where team_id is scraped from the page. Example output for above code::

`['/teams/ATL/2021.html', '/teams/ATL/2020.html']`

`['2020-21', '2019-20']`

3. On each player’s page, you should at least parse the following information:
   a. Basic player’s profile information such as name, position, height/weight, team, birthday (age), recruiting rank, draft team, experience, etc.
   b. Player’s performance statistics.
   c. Player’s salaries.

We loop through all the players who have played in one season and scrape the information to an excel sheet. For example, suppose we are on the 2009-10 season page. On this page we

can see a long list of players who have played in that season. From this page, we parse the player's information such as name, position, age, team and all the player's performance statistics.
Then, we go to each player's individual page and scrape information such as height, weight, age, recruiting rank, draft team.
For the experience of a player, the website has either Career Length or Experience fields on the page. To keep as much as data, we have extracted both these values for the experience of the player. If neither were present, it was given n/a value.
For the salary part, we have extracted the information related to the player's season salary and career level salary.
This scraping is done for all seasons from 2009-10 to 2020-21 in a loop.

4. On each team (franchises) page, at least parse the following information:
   a. Basic team information (top panel on the page) including name, location, seasons played, record, playoff appearance, championship, etc.
   b. Team seasonal statistics.

For each team we scrape the basic data using the following code:

```
details = response.xpath('//*[@id="meta"]/div[2]/p').getall()
```

We save the details in a scrapy item object.
For the seasonal statistics we scrape them from each season page using the following code:

```
for field in self.table_headers:
    value = response.xpath('//comment()').re(r'<tbody><tr ><th.*data-stat="player"
>Team.*<td.*data-stat="'+ field + '" >(.?[a-zA-Z0-9,. ()&amp;]*)</td>')
    team_data[field] = value[0] if value else ""
```

Where the regex allows us to retrieve the information and save it in the scrapy item object.

5. Other information. What other raw data from the site might be helpful to build analytics for team owners? Please include your method, code, data, and reasoning.

To grab the information of all active NBA players and to help us determine the salary of a player, we have scraped each player's advanced statistics found on the player's individual page. Players are paid based on individual performance in respect to numerous on-the-court metrics i.e., points per game, shooting percentage, rebounds, steals, personal fouls, turnovers, blocked shots, and assists. As we had scraped most of the useful information from the player table as mentioned above, we have scraped four additional fields from the advanced table.
These are:

- Player Efficiency Rating - This strives to measure a player's per-minute performance while adjusting for pace.

- True Shooting Percentage - This helps give the measure of team's or a player's actual shooting percentage would be when factoring in free throws and three-pointers (instead of just two-point baskets)
- Win Shares - This helps give an estimate of the wins contributed by a player
- Box Plus/Minus - This will give an estimate of a player's contribution to the team when that player is on the court.

As we wanted a way to help us determine the salary of the player, we scraped these fields in the same csv as player_info.

First we found the advanced table on a player's page

```
advancers =  psoup.find_all('tr', {"id" : advance_text+year})
```

And then we scraped through the fields we wanted as per the advanced table

```
adsoup = BeautifulSoup(str(findallcomments), 'lxml')
per = adsoup.find('td', {"data-stat" : "per"}).text
tsp = adsoup.find('td', {"data-stat" : "ts_pct"}).text
bpm = adsoup.find('td', {"data-stat" : "bpm"}).text
ws = adsoup.find('td', {"data-stat" : "ws"}).text
```

Part B: Data Preparation and Description

Question 1

1. Put all the relevant data variables into one data frame. Explain how you clean your dataset and transform your data variable if any and provide a data dictionary for all your variables.

The data dictionary is added in Appendix A. PSB steps for data cleaning -
We have followed the below steps to clean the dataset and transformation of the variable

Step 1: Column's datatype conversion

1. Data variable Salary is stored as String. So, we converted it to a numeric datatype.
   - Sample salary in the raw dataset - $212,078,086
   - After conversion -212078086
2. Data variables Height and Weight of players are stored as String. They are converted to float data types. Height is converted from feet to inches.
   - Sample height and weight in the raw dataset - 6-5 & 185lb
   - After conversion -77 & 185

Step 2: Handle Missing observations

Columns Rank, Draft, 3P%, FP%, Player_Season_Salary have missing values. Rank is only assigned on a scale of one to one hundred. As a result, the remaining columns can be replaced with 0. 432 players' season salary observations are missing. We discovered that the basketball-reference.com website does not store the salary of players who are earning less than

the minimum salary. As a result, we have dropped such observations. The rest of the rows with null values are deleted. There is no missing value in the team_stats data frame

Step 3: Renaming column names

Column names in season_stats data frame are renamed to make it more readable.

Step 4: Handle duplicates

During any given season, numerous players have played for many teams. Aaron Gordon, for e.g., has played for two teams - ORL and DEN in the 2020-21 season. Therefore, he has 3 observations. Two observations for each team and one combined with Team_ID = TOT.

Player	Year	Season_Year	Draft	Rank	Experience	Height	Weight	Position	Age	Team_ID	Games	GS	Mins_Played	Field_Goal	FG_attempts
Aaron Gordon	2021	2020-21	Orlando Magic	4.0	7	80.0	235.0	PF	25	TOT	50	50	1384	231	499
Aaron Gordon	2021	2020-21	Orlando Magic	4.0	7	80.0	235.0	PF	25	ORL	25	25	736	128	293
Aaron Gordon	2021	2020-21	Orlando Magic	4.0	7	80.0	235.0	PF	25	DEN	25	25	648	103	206

As a result, we are creating two data frames from season_stats. First data frame without Team_ID “TOT” and second data frame with Team_ID “TOT” . All other team observations will be dropped for players who have played for multiple teams in one season..

Step 5: Combine data frame for analysis

We have combined key data variables from both season_stats and team_stats data frames for data analysis where we need data from both DF.

Step 6: Create a balanced season data frame

Created a balanced dataset for balanced panel analysis.

Step 7: Calculate per game statistics for each player in the unbalanced dataset

In order to avoid multiple values scraping we decided to per game statistics using the existing total season data for each player. For attributes that we will be using in further analysis, we are calculating the per-game statistics as total/games.

Question 2

2. For the most current season in the dataset,

a. How many active players are there?

There are 481 players active in season 2020-21

b. How many players are in each position?

The number of players for each position in 2020-21 is
C - 94, PF - 113, PG - 87, SF- 88, SG -115

c. What is the average age, weight, experience, salary in the season?

Average age of players is 25.86 years, weight is 217.84 lb, experience is 5.42 years and salary is $7461400.62

d. What is the average career salary?

Average salary of players in the active season is $38525320.21

Question 3

3. More descriptive statistics on salaries:
a. How many players were active in each season? What is the average salary by season? How about the variance of salary be the season?
There are 434 players in season 2009-10 and 481 players in season 2020-21. Please refer to Appendix B.

b. who are the top 10% best paid players in the most current season? Which teams did these players play for?
48 players are the top best-paid players in the current season. From Stephen Curry earning $43006362 to Malcolm Brogdon earning $20700000. They belong to teams BOS, BRK, CHI, CHO, CLE, etc.

c. who are the bottom 10% best paid players in the most current season? Which teams did these players play for?
The bottom 10% of players are Elijah Bryant earning 24611 to Dewayne Dedmon earning 580811 dollars. They belong to teams like BRK, HOU, ORL, etc.

d. who are the middle 50% by salary? Which teams did they play for?
There are 125 middle 50% player. They are those who are earning in between 1663861.0 and 3562178.0
e. over the career of each of the active players in the most current season, how much money was paid to by season?
There are 481 players in the current season, the total money paid to them was 3588933700.0 dollars.

Question 4

4. Team-player statistics:
Please refer to Appendix C

Question 5

5. What other data from the website can you use to explain salary? Produce a table of summary statistics for key variables you shall use in your analysis in the following Part C. Summary statistics should include at least sample average, standard deviation, min/max.

A player's pay in the NBA can be influenced by a variety of factors. Some variables may have a positive impact on salary, while others may have a negative impact or have no impact at all. We will look at how the independent variables listed below explain our dependent variable(Salary of a player per season) - Experience, Height, Weight, Age, Games, Age, Field_Goal,3_Point_FG, 2_Point_FG, Mins_Played_per_game, Total_Rebounds_per_game, Assists_per_game, Steals_per_game, Blocks_per_game, Turnovers_per_game, Personal_Fouls_per_game, Points_per_game. Summary stats is included in the notebook.

Part C: Data Analytics

Clustering

a. Apply “k-means” algorithm.

We are running kmeans clustering based on the player_season_salary attribute. We choose to run only with the player_season_salary attribute as we don’t want the effect of other attributes to influence the clusters being formed. We want to interpret those attributes based on the clusters formed. We apply the kmeans algorithm using the using the following command:

```
km_cluster1 = kmeans(kmeans_attributes, centers = 4, nstart = 10)
```

Where centers is the number of clusters, nstart signifies the number of random initializations of the centroids.
The size of the clusters are the following: 1176 3450 219 686

b. With k-mean clustering, what attributes about the players do you use, as to justify the players’ salaries?

We have chosen the following attributes to analyse for each cluster:

```
[Mins_Played_per_game, eFG.,FT_percent, Total_Rebounds_per_game,
Assists_per_game, Steals_per_game, Blocks_per_game, Turnovers_per_game,
Points_per_game, Personal_Fouls_per_game, PER, TS., BPM, WS,
Player_Season_Salary]
```

We calculate the means for these numeric attributes for each cluster.

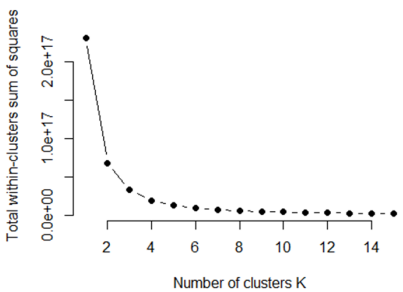
Based on the mean values we can see that the players with the highest salaries on average are the ones who play longer per game and have better overall performance in each category. For example, key metrics like points_per_game, Assists_per_game, Total_Rebounds_per_game, Steals_per_game, Blocks_per_game which are the most integral performance attributes of a basketball player. We can see that for players with higher salaries the averages for these performance attributes are also higher.

We can also see in the advanced statistics like PER (Player efficiency rating), TS% (True shooting percentage), WS (Win shares) and BPM (Box plus/minus) are higher for players with higher salaries.

The means clearly show that players who perform better and contribute more in making their teams win are paid better.

c. How many clusters k do you choose and why?

We plot ‘within-cluster sum of square distance’ as a function of number of clusters. To determine the number of clusters



we have to find the elbow in this plot. Looking at the graph we can select either k=4 or k=3, we select k as 4 as the number of clusters. With k=4 the variances between the salaries in each cluster was lower, hence we slected k=4.

d. Do you get different results with different (random) initialization of centers? Can you find parameters, i.e. k and attributes, that yield stable clustering?

Yes, we get different result with different initialization of centers.

Kmeans clustering algorithm depends heavily on how the centroids are initialization. An example of poor initialization is when when all the N points are close to one of the centroid, in this case the worse case could be K-1 clusters would be empty cluster.

We use nstart to give the number of random initialization kmeans should do. With nstart =10, the algorithm does 10 initializations and chooses the best one with the lowest within-sum-of-cluster sum of squared distances. Also, the default value for the number of iterations (iter.max) is 10. So the algorithm would run till it finds no more changes in the centroid or 10 times.

nstart and iter.max make sure that the result of the cluster is stable and is not entirely dependent on the initialization of the centroid. To prove this, in the code above, when nstart=1 changing the seed value will choose a set of random initial centroids and will give us different cluster sizes in the 2 cases. When the value of nstart is increased to 10 the same cluster sizes are generated.

The parameters for which are getting stable clustering are:

k=4, iter.max=10, nstart=10

e. How do you interpret the clusters to your boss? What do each cluster mean (in terms of the attributes)?

These 4 clusters are grouping of players based on their salaries. With the centers being the average salary of each group.

As mentioned in question-b they allow us to analyze performance attributes and statistics of players once we have the grouping of players based on salary. We can show that better performing players are also the ones who get higher salaries on average. By grouping the players based on salary we can get an insight into traits which are similar between players.

Linear Regression

In the section, we perform linear regression on the NBA dataset to identify the factors that predict a player's salary in the most current season. Apart from the data preprocessing and analysis in part B, following activities were carried out to improve the performance of the model

- We have used the season_unbalanced_per_game_data_final dataset

- We have extracted only the data corresponding to the current active season (2020-21)
- Outlier Removal: There are 82 games in the NBA and scoring 25 points is realistic even for benched players. Hence, we have removed the entries for which the points are less than 25.
- A correlation matrix was plotted to identify the relationship between the dependent and independent variables. It can be observed that out of the variables used Mins_Played_per_game (MPG), Total_Rebounds_per_game (TRPG), Assists_per_game (APG), Steals_per_game(SPG), Turnovers_per_game (TPG), Points_per_game (PPG), Age, Experience, PER, BPM, WS indicate high correlation with the Salary
- After conducting several experiments, we obtained a model with R^2 value of 0.6632 and F statistic score of 85.25. Linear regression was also performed on standardised variables and the RMSE value reduced from 5489000 to 0.587

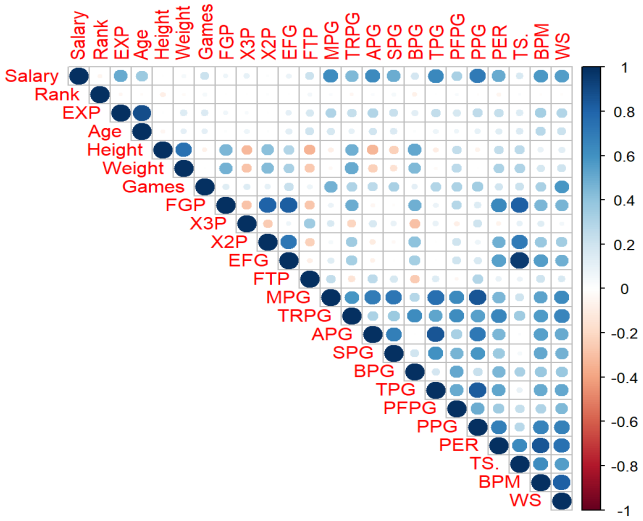


Fig. 1 Correlation matrix between dependent and independent variables

Linear Regression

```
##
## Call:
## lm(formula = Salary ~ ., data = stats_salary_regression)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17894720  -2971290    100771   2467675  21568008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3597475    1816811   1.980  0.04832 *
## EXP           784009      71656  10.941  < 2e-16 ***
## MPG           -545558     113199  -4.819 2.00e-06 ***
## TRPG          1155931     224979   5.138 4.21e-07 ***
## APG           2086325     365450   5.709 2.11e-08 ***
## SPG           2486348    1190904   2.088  0.03740 *
## BPG           2511034     948648   2.647  0.00842 **
## TPG          -3029605     974707  -3.108  0.00201 **
## PFPG         -1268139     602779  -2.104  0.03597 *
## PPG           1431743     162397   8.816  < 2e-16 ***
## PER           -630030     132319  -4.761 2.63e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5489000 on 433 degrees of freedom
## Multiple R-squared:  0.6632, Adjusted R-squared:  0.6554
## F-statistic: 85.25 on 10 and 433 DF,  p-value: < 2.2e-16
```

Fig 2. Summary of Linear Regression Model

```
##          EXP          MPG         TRPG          APG          SPG          BPG
##  0.32785989 -0.47917286  0.28891459  0.42856228  0.09729375  0.11174306
##          TPG         PFPG          PPG          PER
## -0.26517974 -0.09056130  0.97238006 -0.31816409
```

Fig 3. Standardised Beta values from regression model

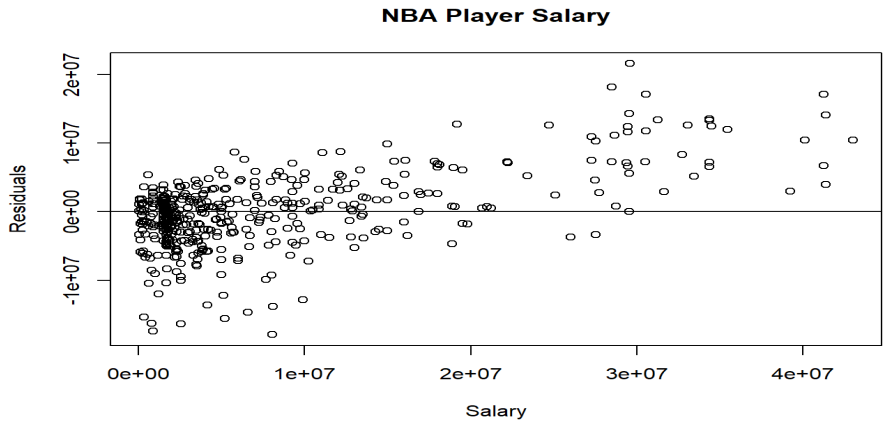


Fig 4. Residual Plot of the player salary

2. Linear Regression. What factors predict a player's salary in the most current season? (10 points)

a. Using salary as the dependent variable, what variables do you include as the independent variables?

Following are the independent variables used in the regression model:
Mins_Played_per_game, TRPG = Total_Rebounds_per_game, APG = Assists_per_game, SPG=Steals_per_game, BPG = Blocks_per_game, TPG = Turnovers_per_game, PFPG = Personal_Fouls_per_game, PPG = Points_per_game, EXP=Experience, PER=Player Efficiency Rating

b. Why do you choose the model you specified? Do you have any theory or rely on your observations?

- It can be observed that the selected independent variables show high positive correlation with the dependent variable Player_Season_Salary
- From experiments conducted, the above combination of variables were giving the maximum R2 value and high F score for the linear regression model and all of them were statistically significant

- Intuitively, the main determinant of a player's salary will be the performance of the player. The variables that we have selected mainly represent the performance statistics of the player.

c. How do you interpret the results, particularly your key variable of interest?

- The model gives an R^2 value of 0.6632 and F statistic score of 85.25.
- We have used standardised beta values and p values to interpret the key variables of interest .
- From the p values, it can be observed that all the independent variables in the model are statistically significant as their respective p values are less than 1% (0.01)
- High value of F (85.25) accompanied by a small value of p less than
- 1% indicates that the null hypothesis is rejected. Hence, we can conclude that there is a relationship between the Player Salary and independent variables.
- Although all the independent variables have an influence on the outcome variable, the variables Points per game (PPG), Assists Per Game (APG) and Experience (EXP) have higher standardised beta values. This indicated that they have higher positive correlation with Player Salary.

d. Which predictors are statistically significant?

All the predictions are statistically significant. The predictors used in the model are Mins_Played_per_game, TRPG = Total_Rebounds_per_game, APG = Assists_per_game, SPG=Steals_per_game, BPG = Blocks_per_game, TPG = Turnovers_per_game, PFPG = Personal_Fouls_per_game, PPG = Points_per_game, EXP=Experience, PER=Player Efficiency Rating)

e. If you are the team coach, what do you tell players from your analysis? For instance, each goal contributes to your salary $X amount, so everyone should shoot as often? How does your analysis contribute to the team's overall revenue model?

- We can determine the influence of each independent variable on the player salary from standardised beta values
- It can be observed that Points Per Game (PPG) has the maximum influence on the Player Salary, followed by Assists Per Game (APG) and Experience (EXP) of the player
- Following are the deductions that can be made by the coach
    - Points per game has a positive correlation with the salary. Players should be encouraged to score more points
    - In addition to scoring more points, coordination among the team members should also be improved by encouraging more assists
    - With experience, the skill of the player improves resulting in higher salary
- From the analysis conducted, we have identified the factors that are important to the player salary. Higher salary is an indication of better performance. Hence, we get a consolidated wisdom about the performance of the players which helps to decide on the

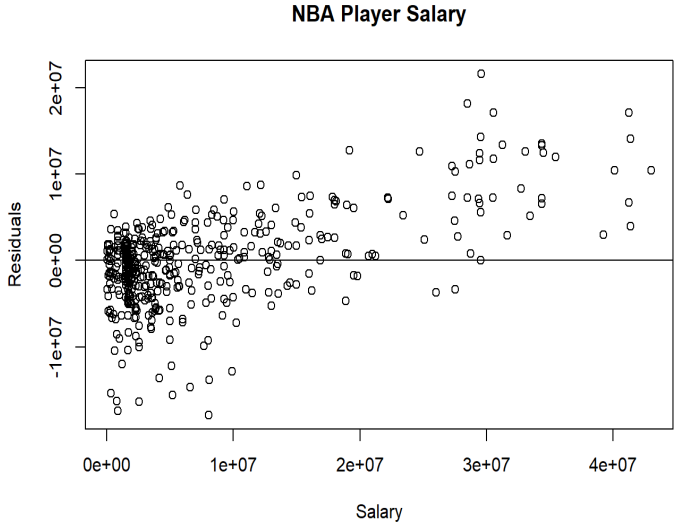
acquisition or retention of the players in the team and thus contributing to the overall revenue of the team.

f. Do you believe the results?

Though the linear regression model shows good adjusted R square value(0.65) and F-statistic (85.25) value we cannot believe it as linear regression model needs to meet the assumption criterias which are mean-zero error,uncorrelated error,linearity, homoskedasticity, normal error, no perfect multicollinearity. To believe the results these assumptions need to be validated which are covered in the next section.

Model Justification

a. Do you think your coefficients in regression are fair, overestimated or underestimated? (hint: check the conditional mean-zero error assumption from the residual plot and what is the implication of the residual average shown in the plot?)

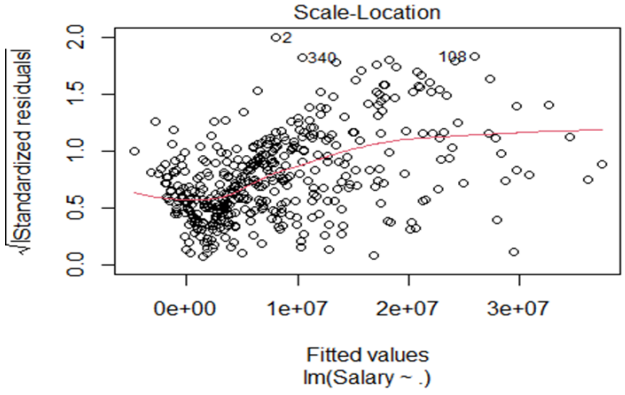


Looking at the residuals plot we can see in the residual plot the data points don't symmetrically scatter around the reference line y=0.

For the lower fitted values the reference line lies above majority of the points. The model appears to be overestimating the values. Whereas, for the higher fitted values the reference line is below majority of the points. For these points it is underestimating the values.

Hence, we can say that the coefficients in the regression are not fair.

b. Do you worry about heteroskedasticity? How can you detect and fix it if any?



We use the scale-location plot to detect heteroskedacity. From the graph we can see the points are not equally spread around a horizontal line.

The variance in residuals is not constant for the lower fitted values.Ideally, the trend should be a flat line, however in our case the

line dips for lower fitted values. This suggests that there is a non-constant variance in the residual error.

To solve for heteroskedacity we can we can apply robust standard errors to fix the standard errors of the model. In the code we use white-huber robust standard error. This will give us standard-errors and their corresponding t-values which are robust to heteroskedacity.

```
##                   Est.     SE_OLS  Est_White   SE_White
## (Intercept)  3597475.2 1816810.83  3597475.2 1373981.65
## EXP           784009.4   71655.48   784009.4   89585.17
## MPG          -545557.7  113198.74  -545557.7  111762.07
## TRPG         1155930.8  224978.55  1155930.8  254694.22
## APG          2086324.9  365449.93  2086324.9  425531.36
## SPG          2486348.4 1190904.40  2486348.4 1375356.15
## BPG          2511034.1  948648.22  2511034.1 1176113.18
## TPG         -3029604.5  974707.32 -3029604.5 1179870.97
## PFPG        -1268139.0  602779.17 -1268139.0  677209.95
## PPG          1431742.5  162396.53  1431742.5  173500.06
## PER          -630029.7  132318.81  -630029.7  108213.09
```

c. Do you worry about multicollinearity among your predictors? How can you quickly tell direction/strength of correlations among your variables?

Yes, Since we have a large number fo attributes in the dataset there is a high chance for multicollinearity.

We can use a correlation matrix, as used above in determining the important attributes to use based on their correlation to each other. It gives a score between -1 to 1 for each attribute with the other attributes. A score of 1 indicates a perfectly positive linear correlation while a score -1 indicates a perfectly negative linear correlation. 0 indicates no linear correlation between the attributes.

d. If there is strong evidence for multicollinearity, which method do you choose to alleviate it and why?

If there is strong evidence of multicollinearity, we can use Principal Component Analysis as a method to allieviate the issue. It is form of dimensionality reduction, which allows us to summarize information of high-dimensional data in a small set of principal predictors.

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6
PC7
## Standard deviation     2.3227 1.2404 0.94689 0.86848 0.67162 0.55641
0.51936
## Proportion of Variance 0.5395 0.1539 0.08966 0.07543 0.04511 0.03096
0.02697
## Cumulative Proportion  0.5395 0.6933 0.78300 0.85843 0.90354 0.93450
0.96147
##                            PC8    PC9    PC10
## Standard deviation     0.50640 0.3193 0.16388
## Proportion of Variance 0.02564 0.0102 0.00269
## Cumulative Proportion  0.98712 0.9973 1.00000
```

Using prcomp on our dataset we can see that the first 3 PCs explain around 78% variability. Hence, we can use them to run our linear regression model.

The model shows similar R2 value as compared to the original model. However, our F-statistic score has significantly increased.

Panel Analysis

In this section, we are using panel analysis to understand what factors predict basketball player's in the past 10 seasons i.e. from 2009-10 to 2020-21.

Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities is observed across time.

Panel data has two dimensions (i,t) to keep track of "entity-time" level data points, denoted by

$$(Y_{it}, \boldsymbol{X}_{it}) \qquad \text{for } n = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T.$$

Panel data can be balanced or unbalanced. In a balanced panel, all panel members have measurements in all periods. If a balanced panel contains N panel members and T periods, the number of observations is n = N×T. Whereas for an unbalanced panel, each panel member in a data set has different numbers of observations. Here, the following strict inequality holds for the number of observations (n) in the dataset: n < N×T.

Over the course of ten seasons, a lot of changes in basketball players' careers may have occurred, resulting in prospective metrics that might be used to estimate their salary. For example, a player may have transferred to another team or been sidelined for a few years due to injury, only to return in subsequent seasons. There could also be considerations relating to playing strategy, such as when a player switches from one playing position to another. In addition, when a player's experience grows over the seasons, his performance may improve, resulting in a raise in salary.

Basketball players' total metrics can be divided into two categories -

- Time invariant variables - Height, Weight, Experience(it is overall experience in our dataset)
- Time invariant variables - Position, Age, Games, Mins_Played, Field_Goal, FG_attempts, FG_percent, 3_Point_FG, 3P_attempts, 3P_percent,2_Point_FG, 2P_attempts, 2P_percent, Free_Throw, FT_attempts, FT_percent, Points, Offensive_Rebounds, Defensive_Rebounds, Total_Rebounds, Assists, Steals, Blocks, Turnovers, Personal_Fouls Player_Career_Salary

Modelling

In this section, we will explore 2 models, namely the Fixed Effect and the Random Effects Model on both types of dataset. For modeling we are going to use -

- Entity variable - player_id
- Time variable - Year
- Dependent variable - Player_Season_Salary
- Independent variable - Height, Weight, Age, Experience, Points, Games, X3_Point_FG, X2_Point_FG, Total_Rebounds_per_game, Assists_per_game, Steals_per_game, Blocks_per_game, Turnovers_per_game, Personal_Fouls_per_game, Field_Goal

Fixed-Effect (FE) Model

The Fixed Effects regression model is used to estimate the effect of intrinsic characteristics of individuals in a panel data set. The “fixed effect” population model has the following form:

$$Yit=t'\delta+Xit'\beta+\alpha i+uit$$

We could apply either first-difference (FD) or time de-mean (FE) as data transformation to eliminate the fixed effect αi. We are not using the First Difference model here because in our case N(player_id) is large compared to T(years) and when N >> T then the Fixed Effect model is preferred. We will apply time de-mean (FE) as data transformation to eliminate the fixed effect. This model accounts for the time-invariant heterogeneity of individuals or time periods(one-way FE model) or both individuals and time periods (two-way FE model).

Random Effect Model

The random-effects model includes the possibility of between entity variations. It also assumes that this variation is random in nature or they are uncorrelated with variables under study.

Hausman test

The criteria to determine “fixed-effect” models (FE or FD) vs. random effect is : if the fixed-effect αi (e.g., here player-level heterogeneity in our dataset) is correlated with covariates in any time period. We will use the Hausman test to choose between a fixed-effects model or a random-effects model. The null hypothesis is that the preferred model is random effects; The alternate hypothesis is that the model has fixed effects.

Panel Analysis on balanced dataset

We have selected players who have played in all seasons to create the balanced dataset. We have 504 observations in which there are 42 unique players spread across 12 years from 2010 to 2021. It also shows that Points_per_game and Total_Rebounds_per_game are highly significant and positively correlated with salary. Hausman test has p value ≤ .05 which means its “significant”, indicating that the fixed effect model is the best fit even for balanced panel dataset.

Panel Unbalanced Data Analysis

We have 5531 observations and 25 variables. In this, there are 1331 unique players spread across 12 years from 2010 to 2021. In this case n < N×T(5531 < 1331×12). A few of the reasons(not all) why this dataset is unbalanced is because players might have retired during this 10 seasons, new players were introduced in mid of this season, players were drafted only in few seasons, Players were out of the game to injury or some personal issue.

Fixed Effect, Random Effect Model, and Hausman Test on unbalanced data set

As per the fixed effect model result, We see that points per game, total rebounds per game, and assists per game are all highly significant, with positive coefficient estimates of 693817.4, 546295.7, and 825533.6, respectively. We can also check that Personal Fouls per Game is highly significant and has a negative impact on a player’s salary, indicating that players who violate the basketball game rule are being negatively impacted in terms of salary. Steals have a negative indicator, which means that while it is an asset to be able to steal the ball from the opponent.

As per the random effect model result, We see that Height, Age, Experience, Points_per_game, Total_Rebounds_per_game, and Assists_per_game are all highly significant and are positively correlated with salary.

The Hausman test yielded a highly significant p-value, indicating that the fixed effect model is the best fit for this unbalanced panel dataset.

Fixed Effect Model	Random Effect Model

```
## Twoways effects Within Model
##
## Call:
## plm(formula = Player_Season_Salary ~ Height + Weight + Experience +
##     Age + Points_per_game + Games + Mins_Played_per_game + X3_Point_FG +
##     X2_Point_FG + Total_Rebounds_per_game + Assists_per_game +
##     Steals_per_game + Blocks_per_game + Turnovers_per_game +
##     Personal_Fouls_per_game + Field_Goal + Year, data = seasons_unbalanced,
##     effect = "twoways", model = "within", index = c("player_id",
##         "Year"))
##
## Unbalanced Panel: n = 1303, T = 1-12, N = 5531
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -17470932  -1578469         0         0   1572808  19372005
##
## Coefficients: (3 dropped because of singularities)
##                            Estimate Std. Error t-value  Pr(>|t|)
## Height                   5786659.4  1704515.3  3.3949 0.0006929 ***
## Weight                  -6929563.6  2104942.4 -3.2920 0.0010028 **
## Points_per_game           693817.4    54982.0 12.6190 < 2.2e-16 ***
## Games                      16224.7     6420.6  2.5270 0.0115418 *
## Mins_Played_per_game      -85495.9    32731.5 -2.6120 0.0090324 **
## X3_Point_FG                -4501.6     3147.1 -1.4304 0.1526811
## X2_Point_FG               -10600.6     1639.3 -6.4664 1.118e-10 ***
## Total_Rebounds_per_game   546295.7    89644.8  6.0940 1.200e-09 ***
## Assists_per_game          825533.6   118355.6  6.9750 3.537e-12 ***
## Steals_per_game         -1508320.0   371617.0 -4.0588 5.022e-05 ***
## Blocks_per_game           114150.7   341242.1  0.3345 0.7380075
## Turnovers_per_game        516167.4   273511.4  1.8872 0.0592039 .
## Personal_Fouls_per_game -1154993.5   216650.1 -5.3311 1.027e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    8.6946e+16
## Residual Sum of Squares: 6.9157e+16
## R-Squared:      0.2046
## Adj. R-Squared: -0.046285
## F-statistic: 83.1819 on 13 and 4204 DF, p-value: < 2.22e-16
```

Hausman Test

```
##
##  Hausman Test
##
## data:  Player_Season_Salary ~ Height + Weight + Experience + Age + Points_per_game +  ...
## chisq = 126.64, df = 13, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = Player_Season_Salary ~ Height + Weight + Age +
##     Experience + Points_per_game + Games + Mins_Played_per_game +
##     X3_Point_FG + X2_Point_FG + Total_Rebounds_per_game + Assists_per_game +
##     Steals_per_game + Blocks_per_game + Turnovers_per_game +
##     Personal_Fouls_per_game + Field_Goal + Year, data = seasons_unbalanced,
##     model = "random", index = c("player_id", "Year"))
##
## Unbalanced Panel: n = 1303, T = 1-12, N = 5531
##
## Effects:
##                        var   std.dev share
## idiosyncratic 1.646e+13 4.057e+06 0.963
## individual    6.367e+11 7.979e+05 0.037
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01879 0.06939 0.11289 0.10379 0.13872 0.17355
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -14963259  -2354820   -188222     10987   1857514  24945580
##
## Coefficients: (1 dropped because of singularities)
##                             Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)             -24747586.4   2298619.7 -10.7663 < 2.2e-16 ***
## Height                     177303.5     33853.6   5.2374 1.629e-07 ***
## Weight                      14346.7      4250.2   3.3755 0.0007368 ***
## Age                        221140.8     21915.9  10.0904 < 2.2e-16 ***
## Experience                 265500.0     24636.4  10.7767 < 2.2e-16 ***
## Points_per_game            578480.6     41209.3  14.0376 < 2.2e-16 ***
## Games                      -13724.3      5022.8  -2.7324 0.0062876 **
## Mins_Played_per_game       -76697.4     21294.1  -3.6018 0.0003160 ***
## X3_Point_FG                  3703.1      2325.3   1.5925 0.1112660
## X2_Point_FG                 -4014.2      1351.3  -2.9706 0.0029724 **
## Total_Rebounds_per_game    487110.5     55726.7   8.7411 < 2.2e-16 ***
## Assists_per_game           753241.6     76166.4   9.8894 < 2.2e-16 ***
## Steals_per_game             17631.6    242767.7   0.0726 0.9421026
## Blocks_per_game            503561.7    213203.1   2.3619 0.0181822 *
## Turnovers_per_game          94038.9    200059.0   0.4701 0.6383153
## Personal_Fouls_per_game  -1115193.6    139127.3  -8.0156 1.096e-15 ***
## Year2011                    4327.3    282973.8   0.0153 0.9877991
## Year2012                 -432001.2    288096.0  -1.4995 0.1337428
## Year2013                 -196860.7    285176.1  -0.6903 0.4899975
## Year2014                 -136665.9    296531.3  -0.4609 0.6448835
## Year2015                 -158850.9    282680.0  -0.5619 0.5741529
## Year2016                  246530.6    285676.9   0.8630 0.3881539
## Year2017                 1538509.0    288047.1   5.3412 9.235e-08 ***
## Year2018                 2186705.7    292391.0   7.4787 7.506e-14 ***
## Year2019                 2471045.9    298259.9   8.2849 < 2.2e-16 ***
## Year2020                 2993110.6    305912.8   9.7842 < 2.2e-16 ***
## Year2021                 3067808.2    314693.9   9.7485 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.9702e+17
## Residual Sum of Squares: 9.612e+16
## R-Squared:      0.51213
## Adj. R-Squared: 0.50983
## Chisq: 6142.91 on 26 DF, p-value: < 2.22e-16
```

Question 4

We have answered the below question on the basis of unbalanced panel dataset analysis.

a. Which time-varying variables matter to explain a player's salary?

The following is a list of time-varying variables that matter to explain a player's salary. In the Fixed effects model for salary prediction, Points_per_game, Total_Rebounds_per_game, and Assists_per_game are all highly significant, with positive coefficient estimates of 693817.4,546295.7, and 825533.6, respectively. We can also see that Personal_Fouls_per_game is highly significant and has a negative impact on a player's salary, indicating that players who violate the basketball game rule are being negatively impacted in terms of salary. Steals_per_game has a negative indicator, which means that while it is an asset to be able to steal the ball from the opponent, it may not be reflected in the player's salary at the end of the game. In the NBA, it's possible that players who steal the ball a lot more are undervalued.

Points_per_game 693817.4 54982.0 12.6190 < 2.2e-16 ***

Games	16224.7	6420.6	2.5270	0.0115418 *
Mins_Played_per_game	-85495.9	32731.5	-2.6120	0.0090324 **
X2_Point_FG	-10600.6	1639.3	-6.4664	1.118e-10 ***
Total_Rebounds_per_game	546295.7	89644.8	6.0940	1.200e-09 ***
Assists_per_game	825533.6	118355.6	6.9750	3.537e-12 ***
Steals_per_game	-1508320.0	371617.0	-4.0588	5.022e-05 ***
Personal_Fouls_per_game	-1154993.5	216650.1	-5.3311	1.027e-07 ***

b. Which are time-invariant predictors in your data set? Do you worry about any fixed effect that might correlate with your predictors? If yes, provide examples of such fixed effects.

“Fixed Effect” αi are time-invariant unobservables. Height, weight, and experience (since our dataset includes overall experience) are time-invariant predictors. Age has also been considered as a time-invariant variable as it is eliminated by the fixed-effect model. Each player might has unobserved heterogeneity αi i.e. traits, characteristics that persistently affect Xit for all t. They are unobserved to us and relatively constant during window of data collection. Such unobserved individual heterogeneity, if unaddressed, will bias the regression result, i.e. biased and inconsistent estimation. I feel that some natural talent of the individual player, as well as experience, may be correlated to other predictors (Experience is captured as time-invariant variable in our dataset). We may successfully solve the potential endogeneity issue from the time-invariant fixed effect by using “fixed-effect” models with panel data. There is, however, a chance that predictors are linked to “idiosyncratic” inaccuracy. Such endogeneity issue and its bias are persistent,and unfortunately, prevailing in many practical cases.

c. Which model of panel analysis do you choose, "fixed-effect" or random effect models and why?

To select the optimal model for our panel study, we used the Hausman test. The null and alternate hypotheses of the Hausman test are listed below -

- Null hypothesis - Random effects is the preferred model.
- Alternate hypothesis - The fixed effect is the preferred model.

The test yielded a highly significant p-value, indicating that the fixed effect model is the best fit for this panel dataset. Accordingly, the FE-model seems to be the most suitable, because we clearly have endogeneity in our model. (The unobserved dependency of other independent variable(s) is called unobserved heterogeneity and the correlation between the independent variable(s) and the error term (i.e. the unobserved independent variable) is called endogeneity.)

d. Based on your results, what do you recommend to the team owner? How much should teams pay existing players?

Based on the findings, we can conclude that players who score more points per game, have good rebounds, and are involved in many assists get paid more. In all of the panel tests we've done so far, these variables have consistently been significant. When determining how much a team should pay existing players, several aspects must be considered. They may assess a player's overall abilities and the worth he brings to the team's overall success. The answers will vary depending on the overall statistics of the squad and the player in question.

Interpretation

How to make sense of your results? Describe the analysis and you are not required to execute the analysis (if you wish, you can). a. Do you think your result could be interpreted in a causal relationship? Why or why not?

No. For a causal effect to exist from variable X to variable Y, 3 conditions must be satisfied

1) Concomitant Variation : X and Y should covary
2) X must precede Y in time
3) Absence of other probable causes . The relationship must be spurious, i.e. it should not be produced by X and Y's joint association with a third variable or a set of variables.

The third condition is the most difficult to satisfy and can only be satisfied through experimental designs which can isolate and/control for the effect of other causal factors. Such a true experiment is designed to maximise the internal validity where we are reasonably sure that the effect (Y) is produced by and only by presumed causal variable which is manipulated (Xs)

Hence, we can conclude that causality can not be strictly established by cross sectional studies. We can only speculate causality provided that we frame the model and hypothesis based on a strong theory (theoretical prediction) and there is replication. Panel studies help us in studying the time element also and hence better than cross sectional studies, but still do not help in ruling out the effect of other possible causes.

b. Winning a game is a team effort, but solely that of a single player. If the objective is to win, how can your prediction include cooperation amongst members?

Given that our objective is to win, there are numerous predictors that can be used to explain team member cooperation. Some of the predictors included by our dataset are a player's experience and the number of assists per game. Experienced players tend to motivate the entire team, and they provide value to the squad by providing valuable feedback on how to enhance the team's overall performance, as they have learned more over the years. Teams gain from the player experience as they try to handle the pressure and retain their cool throughout and some of the most frenetic postseasons. A pass to a teammate that leads directly to a goal is an assist, which is also a form of teamwork. While these are observable predictors, there may be some unobserved heterogeneity. Such as the player's ability to lead, his social motivations, how he communicates with his teammates, and his sportsmanship. They are unobserved by analysts and are relatively constant during the data collection window.

c. New athletes are drafted using a system where teams are randomly set an order (eg.1st for LA Lakers, 2nd for Chicago Bulls, and so on), and get to place a contract with the player in that order. Players then can choose to accept or reject the offer. Which variables do you think matter for a successful draft for the player? Describe what kind of data you need to answer such a question and how.

How actual NBA Draft Works

The NBA allows teams who had worse records (which means that they won fewer games) the previous season to get to pick before better teams. This lets bad teams draft good players and become better.

Random draft order

In order to answer this question, we considered that the draft process is random. Now the word "successful draft" is quite subjective. From the perspective of a drafted player, it might mean a higher salary, being selected by the best team, or being selected by a team where the player can offer value based on his strengths. Again, from the perspective of the team owner, it could mean selecting the best player among all candidates or selecting a player based on the team's present needs.

Based on the foregoing, we believe the following variables might be important for a successful draft in the scenarios listed:

<u>Based on the player's perspective —</u>

For example, a player who excels at collecting rebounds may prefer to play on a team with low total rebound scores. In this way, he can contribute to the overall performance of the squad, and it will be a successful draft for him. Here we need total rebounds stats of both players and all teams participating in the draft.

Another example is a player may want to be part of a team that has a lot of experienced players so that he gets a chance to learn from some iconic NBA stars. For this case, we need the experience stats of all teams participating in the draft.

<u>Based on the team owner’s perspective-</u>

A team, for example, LA Lakers are struggling to score 3 pointers, so for them, a successful draft will be a player who is adept at making them. As a result, the player's 3 point field goal data is required.

Similarly, successful drafting may be accomplished in a variety of ways in the NBA draft process.

Appendix A

Data Dictionary for Raw Data

The raw data is in two sets: **seasons_data.csv** and **team_data.csv**.

1. The seasons_data.csv data set contains data of NBA players' statistics. The set includes 7490 observations and 38 attributes. Each of the columns is defined as follows:
    - Player: Name of the player
    - Season: Season played
    - Draft: Team drafted in. The NBA draft happens every year in June. It is where teams in the National Basketball Association (NBA) choose players who have never played in the NBA before. If a team chooses a player, that player cannot sign a contract to play for any teams other than that team
    - Rank: Rank of the player
    - Experience: Total years of experience
    - Height: Height of the player
    - Weight: Weight of the player
    - Pos: Position (PG Point Guard, SG Shooting Guard, SF Small Forward, PF Power Forward, C Center
    - Age: Age of Player
    - Tm: Team
    - G: Games
    - MP: Minutes Played Per Game
    - FG: Field Goals Per Game
    - FGA: Field Goal Attempts
    - FG%: Field Goal Percentage
    - 3P: 3Point Field Goals
    - 3PA: 3Point Field Goal Attempts
    - 3P%: 3Point Field Goal Percentage
    - 2P: 2Point Field Goals
    - 2PA: 2point Field Goal Attempts
    - 2P%: 2Point Field Goal Percentage
    - eFG%: Effective Field Goal Percentage
        This statistic adjusts for the fact that a 3point field goal is worth one more point than a 2point field goal.)
    - FT: Free Throws
    - FTA: Free Throw Attempts
    - FT%: Free Throw Percentage
    - ORB: Offensive Rebounds
    - DRB: Defensive Rebounds
    - TRB: Total Rebounds
    - AST: Assists
    - STL: Steals
    - BLK: Blocks
    - TOV: Turnovers
    - PF: Personal Fouls
    - PTS: Points per game

- PER: Player Efficiency Rating (A measure of per-minute production standardized such that the league average is 15)
- TS%: True Shooting Percentage (A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws)
- BPM: Box Plus/Minus (A box score estimate of the points per 100 possessions a player contributed above a league-average player, translated to an average team)
- WS: Win Shares(An estimate of the number of wins contributed by a player)
- Player_Career_Salary: Player's career salary
- Player_Season_Salary: Player's salary per season

2. The team_data.csv data set contains data of team statistics. The set includes 360 observations and 44 attributes. Each of the columns is defined as follows:

- team_id: Unique Team ID
- name: Name of the team
- location: Location of the team
- other_names: Other names of the team
- seasons: Total number of seasons played by a team
- seasons_years: First year to last year played by a team
- tot_record: an overall record of the team
- tot_record_pct: record percentage
- playoffs: Total playoffs game played by team
- championships : Total championships played
- season_year:
- g : Games
- mp : Minutes Played Per Game
- fg : Field Goals Per Game
- fga : Field Goal Attempts Per Game
- fg_pct : Field Goal Percentage
- fg3 : Team's 3Point Field Goals
- fg3a : Team's 3 Point Field Goals attempts
- fg3_pct : Percent of Team's 3 Point Field Goals Attempted
- fg2 : Team's 2Point Field Goals
- fg2a : Team's 2Point Field Goals attempts
- fg2_pct : Team's 2Point Field Goals percentage
- ft : Team's Free throw
- fta : Free throw attempt
- orb : Offensive Rebounds
- drb : Defensive Rebounds
- trb : Total Rebounds
- ast : Team's Assists
- stl : Team's Steals
- blk : Team's blocks
- tov : Turnovers
- pf : Personal fouls
- pts : Points per game
- wins : Total games won by team
- losses : Total games lost by team
- off_rtg : Offensive Rating
- def_rtg : Defensive Rating
- arena_name : Arena's name

- attendance : The total attendance per season
- total_salary : Total salary of team
- avg_player_salary : Average salary per team per season
- avg_team_age : Average player's age per team per season
- avg_team_exp : Average player's experience per team per season

Appendix B

a. How many players were active in each season? What is the average salary by season? How about variance of salary be season?

Season Year	Number of active players
2009-10	434
2010-11	445
2011-12	447
2012-13	449
2013-14	385
2014-15	485
2015-16	476
2016-17	484
2017-18	483
2018-19	482
2019-20	480
2020-21	481

```
Average Salary per season
             Player_Season_Salary
Season_Year
2009-10              4.542692e+06
2010-11              1.598152e+06
2011-12              1.082372e+06
2012-13              1.279599e+06
2013-14              2.121178e+06
2014-15              1.079574e+06
2015-16              1.363413e+06
2016-17              1.281229e+06
2017-18              1.681998e+06
2018-19              1.609515e+06
2019-20              1.835243e+06
2020-21              2.206113e+06
Variance of Salary per season
             Player_Season_Salary
Season_Year
2009-10              2.235539e+13
2010-11              5.318920e+12
2011-12              1.204142e+12
2012-13              1.297810e+12
2013-14              1.378882e+12
2014-15              1.393054e+12
2015-16              1.579053e+12
2016-17              1.707684e+12
2017-18              3.475538e+12
2018-19              2.828833e+12
2019-20              4.068773e+12
2020-21              4.384552e+12
```

Appendix C

b. what is the average age of the players by season? Average and variance of experience by season of each team?

Average Team age

Average age of each team by season

name	avg_team_age
Atlanta Hawks	26.757500
Boston Celtics	26.150000
Brooklyn Nets	26.489167
Charlotte Hornets	25.915000
Chicago Bulls	26.497500
Cleveland Cavaliers	26.676667
Dallas Mavericks	27.987500
Denver Nuggets	25.766667
Detroit Pistons	26.160000
Golden State Warriors	26.400833
Houston Rockets	26.425833
Indiana Pacers	26.202500
Los Angeles Clippers	27.717500
Los Angeles Lakers	27.265000
Memphis Grizzlies	26.145833
Miami Heat	28.087500
Milwaukee Bucks	26.400000
Minnesota Timberwolves	25.585000
New Orleans Pelicans	26.045000
New York Knicks	26.628333
Oklahoma City Thunder	25.698333
Orlando Magic	25.782500
Philadelphia 76ers	25.318333
Phoenix Suns	26.045833
Portland Trail Blazers	25.462500
Sacramento Kings	25.569167
San Antonio Spurs	27.846667
Toronto Raptors	25.831667
Utah Jazz	25.777500
Washington Wizards	26.540833

Age Team experience

Average experience of each team by season

name	avg_team_exp
Atlanta Hawks	4.892500
Boston Celtics	4.525000
Brooklyn Nets	4.640000
Charlotte Hornets	4.444167
Chicago Bulls	4.457500
Cleveland Cavaliers	4.822500
Dallas Mavericks	5.894167
Denver Nuggets	4.333333
Detroit Pistons	4.401667
Golden State Warriors	4.684167
Houston Rockets	4.400000
Indiana Pacers	4.252500
Los Angeles Clippers	5.915833
Los Angeles Lakers	5.789167
Memphis Grizzlies	4.145833
Miami Heat	6.220833
Milwaukee Bucks	4.579167
Minnesota Timberwolves	3.817500
New Orleans Pelicans	3.914167
New York Knicks	4.873333
Oklahoma City Thunder	4.265833
Orlando Magic	4.230000
Philadelphia 76ers	3.588333
Phoenix Suns	4.110833
Portland Trail Blazers	4.087500
Sacramento Kings	3.853333
San Antonio Spurs	5.695000
Toronto Raptors	4.100000
Utah Jazz	3.845000
Washington Wizards	4.725000

Team experience Variance

Variance of experience of team by season

name	avg_team_exp
Atlanta Hawks	1.927257
Boston Celtics	2.590627
Brooklyn Nets	1.199127
Charlotte Hornets	0.722172
Chicago Bulls	1.944475
Cleveland Cavaliers	3.553711
Dallas Mavericks	2.653154
Denver Nuggets	0.643733
Detroit Pistons	0.478961
Golden State Warriors	1.795445
Houston Rockets	2.072618
Indiana Pacers	0.519511
Los Angeles Clippers	1.909172
Los Angeles Lakers	1.730154
Memphis Grizzlies	1.220554
Miami Heat	1.882172
Milwaukee Bucks	1.155536
Minnesota Timberwolves	0.601911
New Orleans Pelicans	0.863190
New York Knicks	1.894788
Oklahoma City Thunder	0.853681
Orlando Magic	1.735164
Philadelphia 76ers	1.253724
Phoenix Suns	1.337772
Portland Trail Blazers	0.607020
Sacramento Kings	0.701770
San Antonio Spurs	1.353809
Toronto Raptors	0.308382
Utah Jazz	1.053773
Washington Wizards	1.492391