Final Report for MA678 Project

Daniel(Chen) Xu BUID:U49903384

12/10/2021

Abstract:

The National Basketball Association (NBA) is a professional basketball league in North America. The league is composed of 30 teams (29 in the United States and 1 in Canada) and is one of the four major professional sports leagues in the United States and Canada. It is the premier men's professional basketball league in the world. The NBA regular season runs from October to April, with each team playing 82 games. As the premier professional basketball league in the world, the change in style of basketball in the NBA led to the change in style of modern basketball. Here rises problems: Which statistical data of basketball best captures the changing style of modern basketball and how these types of data determine the injury situation for players of each season in today's style of basketball? To address those problems. I use Exploratory Data Analysis to find the factors related to the change of play style and use some statistics to build a multilevel model. However, the model shows that the variables all have a slight impact on the injury situation for each player and is slightly different between seasons as the style of basketball changes as time passes. This report consisted of 6 main parts: Abstract, Introduction, Method, Result and Discussion.

Introduction:

The game of basketball has worldwide appeal. It requires speed, athleticism, skill and the ability to stay calm in the most hectic moments of the game. Basketball has undergone many changes since James Naismith invented the game to give his students something to do when the cold weather prevented them from playing sports outside. As the representative basketball league in the world, in order to gradually improve the ornamental and competitive basketball. The NBA has gone through several landmark rule changes. These rule changes have directly or indirectly affected the development and style of basketball. For example, the establishment of the 3-point line, the establishment of the defensive 3-second violation, the expansion of the 3-second zone, etc. From my point of view, two changes contributed to the style of basketball now, the first one is setting the hand check as a blocking foul outside the 3-point line and the second is the establishment of the 3-point line. The hand check is a defensive maneuver that is made on the ball-handler that a defender would extend an arm and use his hand to initiate contact with the ball-handler. Without defenders using their hands to initiate contact outside the 3-point line, players nowadays have more space to take a relatively more efficient offensive choice: shoot 3-point. I will present how the average number of 3 point attempts per game change from the Season 2010 - 2011 to the Season 2019 - 2020 in my EDA part. When it comes to professional sports, injuries are a perennial problem. For an NBA player, the degree of injury in one season directly determines his performance in this season or even the whole career. In this report, I will analyze that in the current style of basketball, what factors affect the average injuries level of NBA professional players each season.

Therefore I use a multilevel model to see what and how factors may influence the average injuries level of a player in a season. Before that, I clean the data and combine some information collected from Kaggle. I will summarize my process for data cleaning in the following part.

Method

Data Cleaning and Processing:

The main data set is published on Kaggle: Basketball Players Stats per Season - 49 Leagues includes 11,000 players details and stats per Season from the 1999-2000 season through the 2019-2020 season. And in order to analyze the injury situation for players, I also found a data set on Kaggle: NBA Injuries from 2010-2020 includes detail on every injury in the NBA from the beginning of the 2010-2011 season through the end of the 2019-2020 season.

To clean the stats dataset:

Firstly, I subset the original dataset by only selecting players who played for NBA from the season 2010-2011 to the season 2019-2020; Secondly, I created a new variable age to store every player's age at that time because I think age is one of important factors that have effect on the average injuries level for players and it can be a factor to distinguish same player from different seasons; Thirdly, I calculate average stats for every players for each season and store them in a new dataset because from my perspective, using average data as factors can more intuitively show how the players' court performance will change with the change of basketball style. Using aggregate data is too general, and if these averages are a significant factor in the season's injury levels, then more detailed adjustments can be made to players' performance to prevent potential injuries; Finally, I separate my dataset into Regular Season and Playoffs and paste season and players name to a new variable join_c to prepare for joining datasets. I will only focus on the regular season.

To clean the injuries dataset:

Firstly, I classify each injury by body part; Secondly, I standardize dates and summarize all the dates into 10 seasons; Thirdly, paste season and players name columns to a new variable join_c to prepare for the joining of datasets.

To wrangle datasets for further analysis: After I cleaned two datasets, I left join them together by the variable: join_c and select the variables that I need for further analysis. And I also assigned levels for injuries according to the standard estimated recovery time and calculated average injury level for each player by generate the mean levels for injuries each season. Below is the table for injury level corresponding to each injury body part:

Injury body part	Injury level	Injury body part	Injury level
None	1	Torso	11
Rest	2	Back	12
Illness	3	Leg	13
Finger	4	Foot	14
Hand	5	Toe	15
Face	6	Ankle	16
Arm	7	Groin	17
Shoulder	8	Hip	18
Neck	9	Knee	19
Head	10	Achilles	20

Then, the final version of my cleaned dataset has 2399 observations and 30 variables. I will only choose the variables I listed above for my further analysis.

Exploratory Data Analysis:

For the first step of my exploratory data analysis, I present the change of number of 3 point attempt per game for each players from the season 2010-2011 to the season 2019-2020. In order to show the trend of change more intuitively, I use violin and boxplot to show the distribution per season and use an arrow to connect each mean point to show the changing trend. We can conclude from the boxplot in Figure. 1, the number of three-point attempts is steadily increasing and there is a clear trend of growth start from the season 2014-2015. And there's a steady stream of players averaging more than 10 three-point attempts per

game in a season. For the violin plot, we can conclude that more and more players began to attempt the 3-point shot in the game and gradually made the 3-point shot a regular scoring option. Now we know that shooting more three-pointers is a game changing trend for today's basketball game. However, how this factor affect the average injuries level for a player? Let's explore more plot to select factors to build my multilevel model.

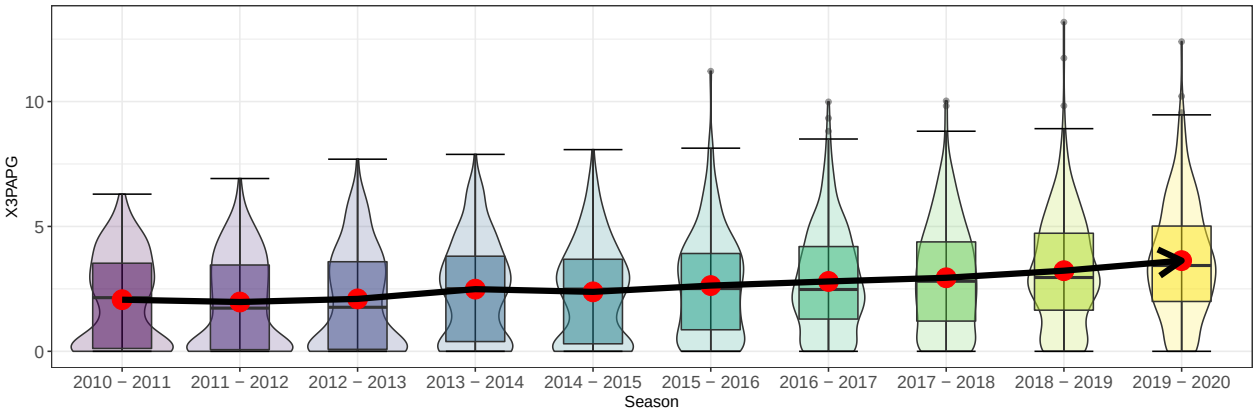


Figure 1: Number of 3 Points attempt per game distribution from season 2010-2011 to season 2019-2020

Because of the length of the report, I only picked four factors: `MPG`,`GP`,`X3PAPG` and `PFPG` to display their correlation with the average injuries level. I will use table to show the reason I choose those four factors:

Variable Names	Explanation	Reason
MPG	Average minutes a player play per game	Shows a player's fatigue per game
GP	The number of games a player play in a season	Shows a player's fatigue per season
X3PAPG	Average number of 3 points attempt per game	The on-court statistics that best represent the evolution of modern basketball
PFPG	Average number of personal fouls per game	The contact during play time that is most likely to injure a player

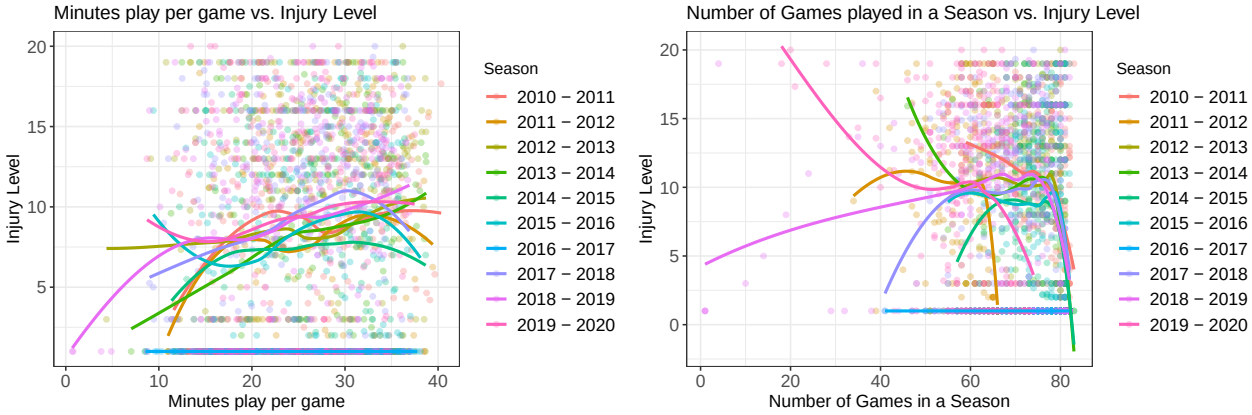


Figure 2: Correlation between the playing time per season and the average injury level

From Figure.2, although the relationships are not linear, we can conclude that as a player's fatigue index

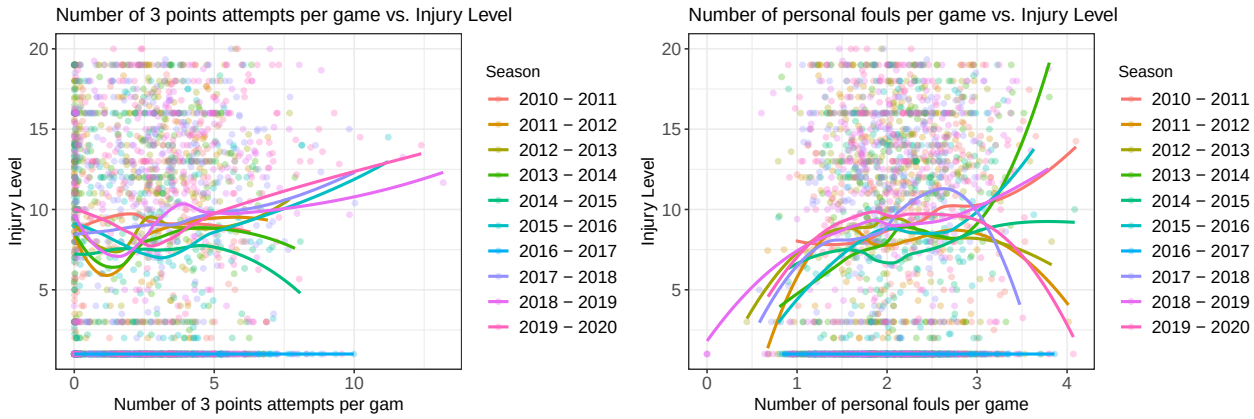


Figure 3: Correlation between some of the avaerga stats per season and the average injury level

increases, his likelihood of serious injury increases.

From Figure.3, although the relationships are not linear, in a certain range, we can conclude that the average injury index has a positive impact with the number of 3-point attempts and the number of fouls. In other words, the more three-point attempts and personal fouls a player makes within a certain range, the more likely he is to suffer injuries.

Model Fitting:

Here are explanations of variables I selected for model building:

Variable Names	Explanation
Season	NBA regular season year range
Age	Player's Age
PER	Player Efficiency Rating
GP	The number of games a player play in a season
MPG	Average minutes a player play per game
PFPG	Average number of personal fouls per game
X3PAPG	Average number of 3 points attempt per game
height_cm	Player's height in centimeter
weight_kg	Player's weight in kilogram
injury_level	A level summarizes the average injuries situation for each player of each season

Considering different regular seasons, I will use multilevel model to fit the data. From EDA above, it is clear that for different seasons there are different correlations with variables I selected, so I use varying slope and varying intercept in multilevel models. Besides, just as I mentioned before, regular season will only be taken into account because I found that in the playoffs, players are more engaged and focused on recovery than they were during the regular season. And the playoff schedule is not as dense as the regular season. The players are rested and ready for every game. As a result, a majority of the injury level for playoffs subset equal to 1, which means no injury. Below is the function:

$$lmer(injury_level\ height_cm + weight_kg + Age + MPG + X3PAPG + GP + PFPG + PER + (1|Season))$$

And to see the fixed effects below, only the variable MPG is significant at alpha = 0.05 level, I use a table to show all the coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.931	4.113	2.415	1.574e-02 ***
height_cm	0.014	0.025	0.547	5.843e-01 ***
weight_kg	0.013	0.019	0.700	4.837e-01 ***
Age	-0.002	0.029	-0.056	9.554e-01 ***
MPG	0.141	0.026	5.376	7.611e-08 ***
X3PAPG	0.0413	0.074	0.561	5.750e-01 ***
GP	-0.156	0.015	-10.700	0.000e+00 ***
PFPG	0.375	0.258	1.454	1.460e-01 ***
PER	0.025	0.030	0.820	4.125e-01 ***

Result

Model Coefficients interpretation

In order to interpret my model, I take an example here, for the season 2019-2020, we can conclude this formula:

$$InjuryLevel = 9.82 + 0.014 \cdot height_{cm} + 0.013 \cdot weight_{kg} - 0.0016 \cdot Age$$
$$+0.14 \cdot MPG + 0.041 \cdot X3PAPG - 0.16 \cdot GP + 0.38 \cdot PFPG + 0.025 \cdot PER$$

Because generally speaking, injury is something that is difficult for a professional athlete to avoid, so it makes sense that intercept in the formula is positive. And 9.82 is approximately the mid value of my injury level. And among 8 indicators, height, weight, minutes play per game, 3 points attempts per game, personal foul per game and player efficiency rating all have postive slopes, which means they all have positive impact on the average injuries level for each player. For each one unit difference in minutes play per game, the predicted difference in the average injuries level is 14% with other indicators keep constant. In other words, if a player play for one more minutes per game in the season 2019-2020, his average injuries level will increase by 14%.

And we can see from the table above, there is only one factor (minutes play per game) has significant effect to the average injuries level, which means that in the model I build, height, 3 points attempts per game, personal foul per game and player efficiency rating are not main effect to have influence on the average injuries level. I will show some plots for model validation in appendix.

Discussion:

The estimates are reasonable in some extents. From the player's individual body statistics, when a player with higher height, heavier weight, he would suffer some injury that need more time for recovery. And more 3-pointers a player shoot per season, the higher level of average injury the player will have. In other words, from my model, players are more likely to have injuries that needed longer time period for recovery in nowadays NBA games. My guess is, with today's style of basketball. With the increasing number of three-point shots, the pace of the game is getting faster as players can choose the relatively efficient means of scoring without having to go round after round to find a mid-range shot or attack to the basket. This increases the average number of runs a player can run and increases the likelihood of serious injuries. And the reason for other two positive affect factores player efficiency rating and personal fouls per game, my explanation is that the higher the efficiency of a player's game, the more it can reflect the player's abilities. Then the defensive pressure on the field of this player will be much higher than that of other players, which increases the possibility of serious injuries. In the same way, when a player averages more fouls per game, it means that his defensive enthusiasm will be relatively higher, which will also increase the possibility of serious injuries.

The result of model almost matches well for different categories in EDA part, but there are still some problems. In EDA part we see that the regression lines are not all positive trend or negative trend. And for the season 2016-2017, the regression line is a horizontal line, which means there is some problems with injury data collected. The injury number is much smaller when compared to other seasons.

I only used eight variables to predict the outcome, but there are still some variables that may have big effects on the situation of injuries. And because the human body is involved, there will be many unstable factors. The data that can be collected from basketball games is very large. And for my outcome: injury level, I only assign the level according to the recovery time, the definition is a bit vague, which leads to my model not being able to locate useful factors efficiently. For further improvement, I will adjust the definition of my outcome and include more data in my analysis.

Citation:

1. https://www.sportsrec.com/358083-the-history-of-basketball-for-kids.html
2. https://dunkorthree.com/hand-checking-rule-nba/
3. https://www.kaggle.com/jacobbaruch/basketball-players-stats-per-season-49-leagues
4. https://www.kaggle.com/ghopkins/nba-injuries-2010-2018

Appendix

Model Validation:

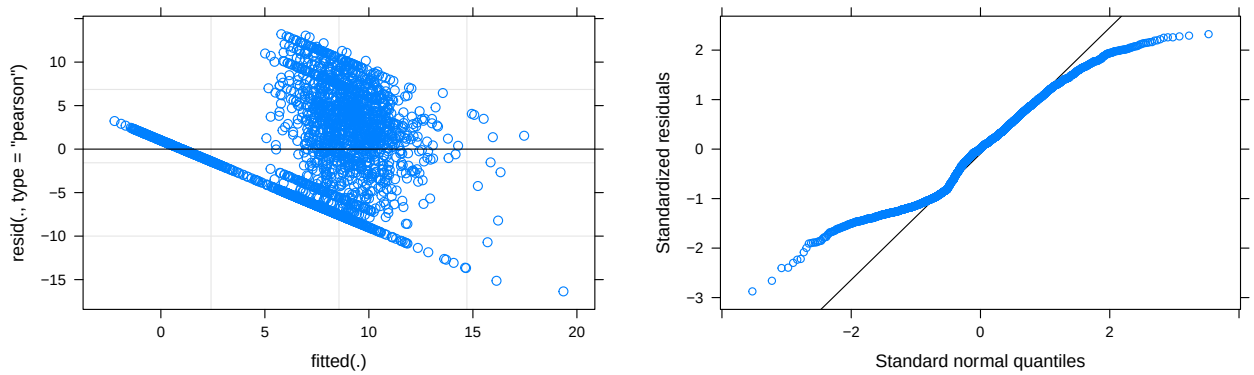


Figure 4: Residual plot and Q-Q plot.

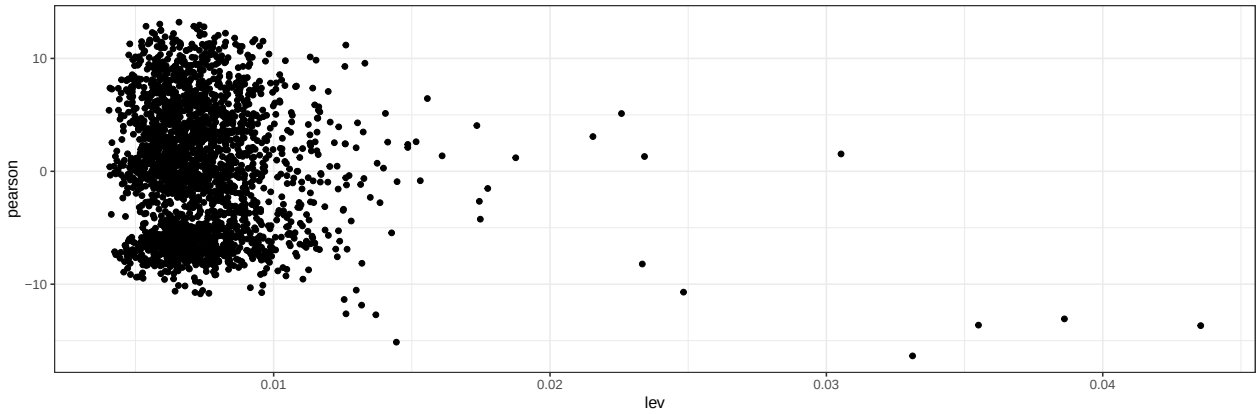
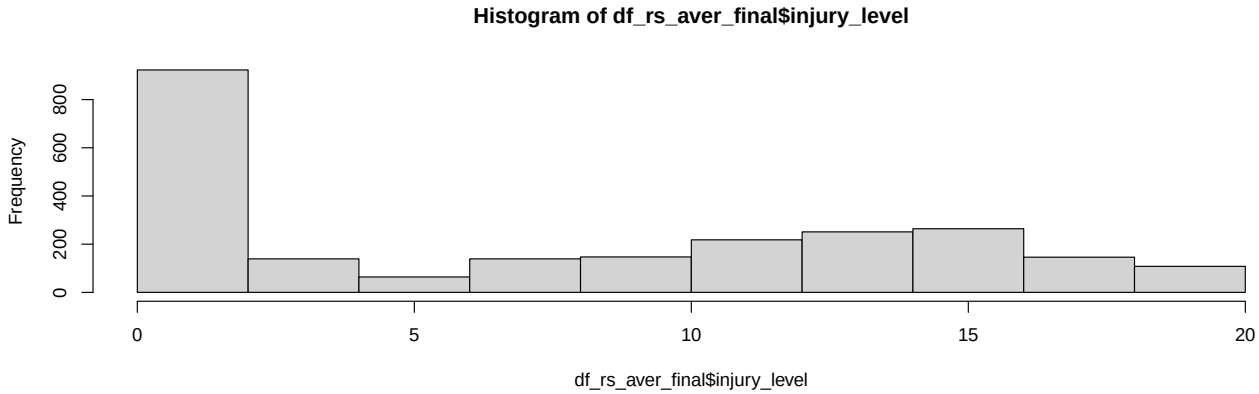
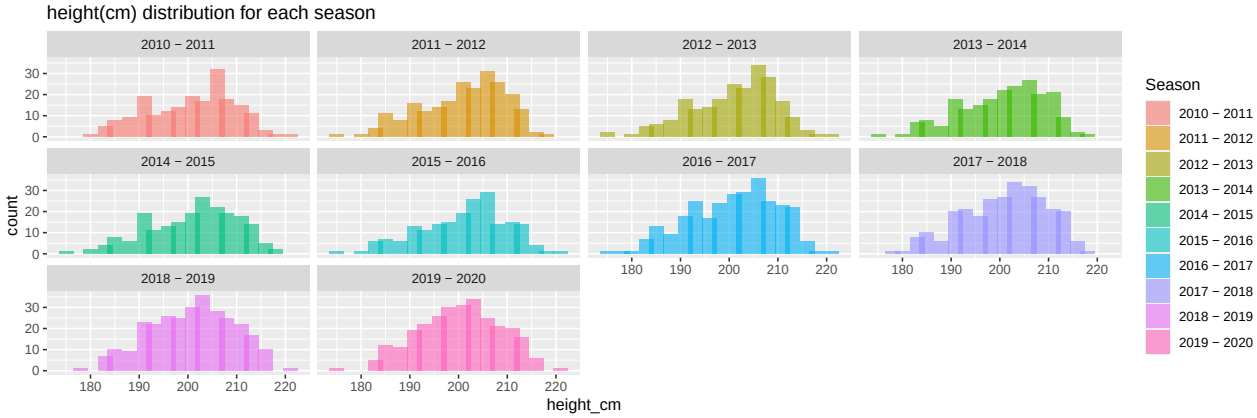
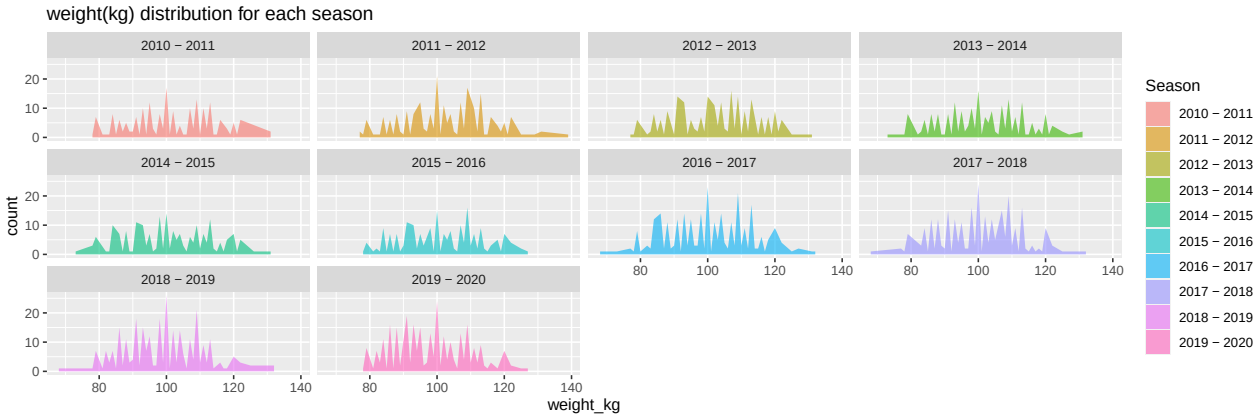


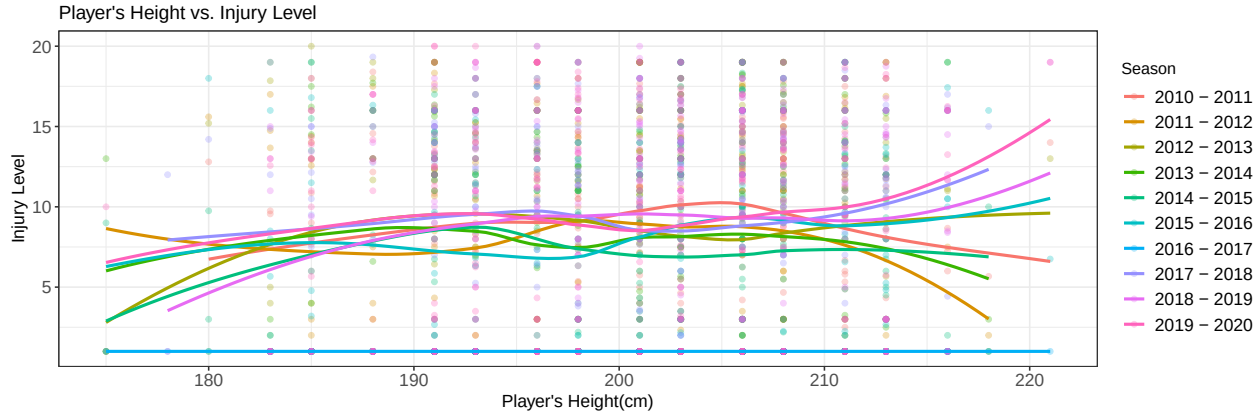
Figure 5: Residuals vs Leverage.

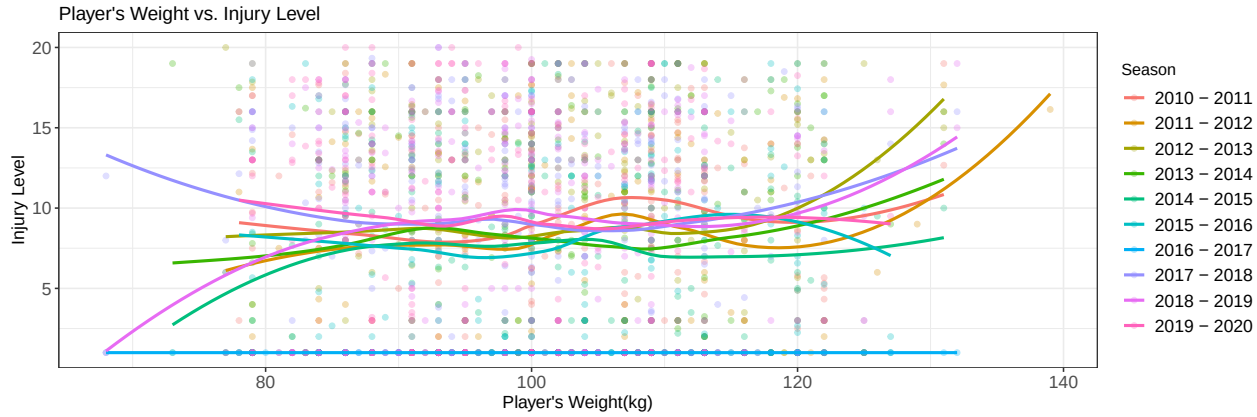
More EDA

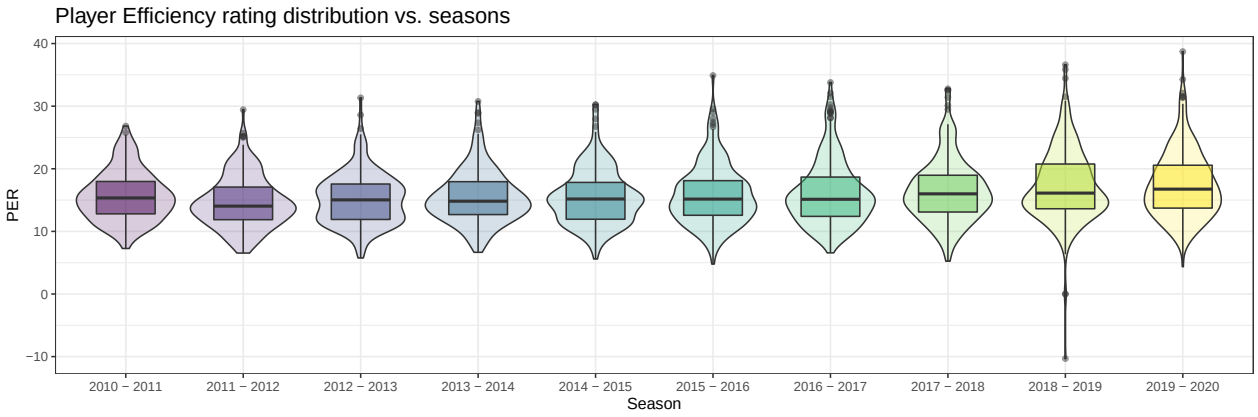



height(cm) distribution for each season
2010 – 2011
2011 – 2012
2012 – 2013
2013 – 2014
2014 – 2015
2015 – 2016
2016 – 2017
2017 – 2018
2018 – 2019
2019 – 2020
count
height_cm
Season


weight(kg) distribution for each season
2010 – 2011
2011 – 2012
2012 – 2013
2013 – 2014
2014 – 2015
2015 – 2016
2016 – 2017
2017 – 2018
2018 – 2019
2019 – 2020
count
weight_kg
Season


Player's Height vs. Injury Level
Injury Level
Player's Height(cm)
Season
2010 – 2011
2011 – 2012
2012 – 2013
2013 – 2014
2014 – 2015
2015 – 2016
2016 – 2017
2017 – 2018
2018 – 2019
2019 – 2020


Player's Weight vs. Injury Level
Injury Level
Player's Weight(kg)
Season
2010 – 2011
2011 – 2012
2012 – 2013
2013 – 2014
2014 – 2015
2015 – 2016
2016 – 2017
2017 – 2018
2018 – 2019
2019 – 2020


Player Efficiency rating distribution vs. seasons
PER
40
30
20
10
0
−10
2010 − 2011
2011 − 2012
2012 − 2013
2013 − 2014
2014 − 2015
2015 − 2016
2016 − 2017
2017 − 2018
2018 − 2019
2019 − 2020
Season