5-2021

Sports Data Analysis – Application of Sports Data In Athletics

Zhong Zhuang

Sports Data Analysis –
Application of Sports Data In Athletics

Zhong Zhuang

Advisor: Richard Aroian

School Of Professional Studies

May 2021

Contents

Abstract

With the data technology developing, Sports industry beginning to more and more pay attention to data application in sports business and athletics. Especially in sports athletics field, sports professional teams managers take advantages of data to draft new players and analysis games. Recently, in sports history, some famous sports cases involve sports data analysis like MLB Oakland Athletics “Money Ball”. This is research goal is focusing on how data be applied to sports athletics field.

The target readers are some peoples who are love sports game and want to know how the data be used in sports athletics. This research can help them understanding what is meaning of sports data. In this research, I will try to use easy understanding words to explain the data and analytics methods in sports. The research methods include reading related books and searching related content by internet (qualitative analysis), quoting practical stories about sports data applications in real sports world (case study), and summarizing key points about sports data application literatures.

The results demonstrate the importance of sports data application and broad development prospects. Based on research, proved the hypothesis. At the same time, the research results including describing and summarizing sports data, statistical methods, and modeling relationships using Linear regression. The research also includes some classical stories about sports data application in real world for benefits to understanding the data applications in real cases.

Chapter 1

1.a General Introduction of the Research Project

In 2011, a sports movie – Moneyball had been released. This movie talks about how a small ball market professional baseball team manager and his assistant took advantage of mathematics and data to make up a baseball team and create a competitive season with minimal budget. This movie really inspire me how the sports data be used in real professional sports games. Because this movie breaks traditional logical and advice of based recruitment. It completely based on data analysis. Data has become a hot point in modern time. You can easily find many industries need data to support them to make better decisions for group successful in some special fields. The Sports industries has also begun to study the importance of data application. Especially, in sports athletics area, when we watching any sports games, we always watching and meeting some data about a players or team's key data display. But, if nobody to interpret these data, these data will become no valuable. These data are benefit to us to better understanding the game and player performances. The same reason, with sports development of the sports industry and the investment of money, the managers hope to build a competitive team and attract more fans with good performances. And then, the team can make much money from fans and supporters. So managers want to use data to make good decision to choose ideal player for them teams. After 2000 years, the practical

application of sports data began to become more and more famous. Let me take some instances, Oakland Athletics baseball team manager Billy Beane, Houston Rocket basketball team manager Daryl Morey, and NHL Chicago Blackhawks. Daryl Morey is big fan for sports data analysis. He graduated from MIT and Northwestern University. As a smart guy, he thought high effective offensive three-point is critical for team successful by his analysis. And then Houston Rocket begun to shot many more three-point shots than before. He built a team that has always been considered strange and he even trade a good big player Clint Capela.

My research about data applications in sports athletics field. The content including many aspects of sports data such as stories, analytics methods, types of sports data… I want to show you feasibility of data applications in Sports and how data is used in the sports analysis. I will introduce some very basic concepts and theories about sports data analysis. According to my research problems, I also raised related problem, what are the development prospects of sports data? From my research, I summary some personal research results about sports data analysis.

The main reference materials and literatures for this paper are Thomas A. SEVERINI's Analytic Methods In Sports, Lorena Martin's Sports Performance Measurement and Analytics, Ben Taylor's Thinking Basketball, And Wayne Winston's Analytics Stories. Some other extra materials will be found in Internet. The main research technology is literatures summaries and case study.

The data applications is becoming popular in sports industry. It breaks a lot of traditional ideas in sports industry. More and more professional teams use data to make advantage in sports athletic and business. But at the same time, you will find some teams still keeping some former professional players as scooting. The first hypothesis of this research about whether data can replace coaches to guide the game. My original hypothesis is the data can replace coaches to guide games. The hypothesis based on a question, If our data analysis enough strong, do we still need some people who own extensive professional sports experiences? At least you can find that many professional team general managers have never participated in any professional athlete experiences. With developing of Artificial intelligence, machine learning are growing stronger than before. Maybe one day, we will see two super computers arranging tactics to guide humans to play games. The second hypothesis of data analysis is meaningful for predicting the outcome of the game. We are seeing some sports games gambling agencies take advantage of data to predict games results. If we using data to predict outcomes of games, it is really reliable? These problems will be explored in this research paper. I will try to summary reading materials content and online materials to find some answers and reasonable results.

The main results of the research is to demonstrate and prove the hypothesis and illustrate the research problem, successfully exploring a field of my interest. My personal contribution is to independently complete all the work of this research paper, including literature reading and collation, problem analysis and document editing.

1.b Rational for Research Project

First of all, many famous sports management event have proved that the application of data in the sports industry is feasible. For example, Oakland A's manager use minimal budget to make the team a competitive team.

Secondly, data is everywhere. We can easily discover the application of data in sports. Whatever in live broadcast of sports game or some sports news software such as ESPN. You can always find kinds of data records and predictions. So data also in sports athletics area. But no everyone has time or enough knowledge background to exploring this sports data secret. This research discussed the professional problem of data analysis from the side. Many data analysts have good data analysis skills, but they are not up to the job of sports data analysis. It is reasonable to explore the field professional data analysis. It has the same rationality as financial data analysis and medical data analysis.

Finally, in the modern world, sports analytics can do so much more. Teams can use data to prevent plays injuries. Coaches can choose the best players to players to play games based on data analysis. But how they use data analysis method behind these behaviors. This is worth to exploring it. This research help sports fans know how the data is processed. For someone who want to explore specific data analysis further, the research help them understand the difference between sports data analysis and general data analysis.

1.c Definition and Explanation of Key Terminology

Sports Analytics: Sports analytics is a method of combine data and statistical analysis. It is also a practice of applying mathematical and statistical principles to sports in specific activity.

Data: data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.

Quantitative analysis: is a technique that uses mathematical and statistical modeling, measurement, and research to understand behavior.

Qualitative analysis: it involves collecting and analyzing non-numerical data to understand concepts, opinions or experiences.

Variables: are any characteristics, number, or quantity that can be measured or counted.

Statistics: is the science concerned with developing and studying methods for collecting analyzing, interpreting and presenting empirical data.

Probability: means possibility. It is a branch of mathematics that deals with the occurrence of a random event. The value is expressed from zero to one. Probability has been introduced in Mathematics to predict how likely events are to happen.

Linear Regression: is a basic and commonly used type of predictive analysis. It can examine does a set of predictor variables do a good job in predicting an outcome (dependent) variable and which variables in particular are significant predictors of

the outcome variable, and in what way do they-indicated by the magnitude and sige of the beta estimate-impact the outcome variable.

Machine learning: is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learns from data, identify patterns and make decisions with minimal human intervention.

Sample: refers to smaller, manageable version of a larger group. It is a subset containing the characteristics of a larger population.

Chapter 2

2.a Brief Overview of Theoretical Foundations Utilized in the Research Study

Because we focus on data analysis, so theoretical foundation is statistical concepts. Statistical concepts are central to methods presentation. Statistical methodology is a vast topic, fortunately, there are a few central concepts and basic methods that can greatly improve our understanding of data and the processes that generated them. Statistic is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting raw data

In sports data method analysis, statistics play at least two important roles. One is the use of statistical methodology to efficiently extract relevant information about measurements and their relationships. Statistical models are essential theoretical in this research process. When we using some statistical model, we can find some general features of relationships between variables. These statistical model can be used in many other fields and as core of statistical methodology. Statistical theoretical help us find meaningful relationships in data. The second role of statistical concepts in analytic method is to provide a framework for using probability to describe uncertainty. Because sports game results have random nature, so any conclusions we draw from analyzing sports data will naturally have some uncertainty and express it. Recognizing random nature is main contribution of analytic method

for sports data analysis. Some basic statistic method including Mean, Median, Standard Deviation(SD), Interquartile range(IQR). Mathematic also involve in this research.

In the real world, we have to do analyses base on available data. So statistical methods play two basic roles under this situation. The first role is statistical provides methods for extracting the maximum amount information from a set of data. The second role is statistical provides us a way to quantify the uncertainty that results from having to base these conclusion on such limited data.

Another Theoretical Foundations is Probability. Probability concerning numerical descriptions of how likely an event is to occur. As we all know, the reason why do we want to use analytics method because we want to use data to better understand the factors that influence the result of sporting events. But we must be noticed all of sports facts have a random element that come into play. So we need to understanding this randomness is crucial to being able to make conclusions from sports data. Probability theory is the branch of mathematics that deals with random outcomes. In this research, we need to pay attention to basic properties and rules of probability that used in sports data analytics. Some appropriate probability theory is the concept of experiment of generates a random outcomes(events). When we have some data collections, we can use probability to predict whether some events would happened. If A is a possible event in an experiment, we normally denote the probability of A by P(A). Probability can describe any possible specific outcome that might occur when an experiment is performed. But in sports data analysis, we

generally concern about data rather than events. So random variables provide the mathematical link between probability theory and data. The random variable is derived from the outcome of an experiment as a numerical quantity. And the random variable can be used to define events. Normally, we write it as P(X=x), X is an event, x demotes a possible value of X. Depending on probability theory, the set of values P(X=x) for all possible x is called the probability distribution of the random variable.

This research involved statistical and probability, some very basic problems need mathematic as fundamental theories. The basis mathematical method including addition, subtraction, multiplication, and division. Depending on our research problem need to make conclusion from our reading and literature materials. So we do need to take advantages of mathematical methods to help us make summary and conclusion from reading. One of methods is Induction. This is not mathematical induction, but the theory is same. The process of mathematical induction theory logical is like playing dominoes. If you want to knock down all the dominoes, first you must knock down the first card, and then you must ensure that every card can knock down the next one, then all the cards will be down. For example, let P(n) is a statement that depends on the natural number n, and M is another natural number. Then if P(M) is true, whenever P (k) is true for k >= M ,then P(k+1) is also true, then P(n) must be true for every number n >= M. The same theory logical can be used in literature summary. So we need to recognize something must true, and then some other conditions can be added as relevance, which must also be true. Depending on

this theory logical, we need to make conclusion about some core ideas and view them is true, and then add some necessary other ideas, we can get a new idea.

The linear regression is another method under foundation of mathematics theory. In linear regression, we can determine that best-fitting line cross the data and use it to better understand the relationship between the variables under consideration. The linear regression model can be used in machine learning. We can collect necessary enough data and put them into linear regression model and then training our model. After we get a line from data, we can use this model to predict other variable results. In this research, we need to understanding about linear regression logical, because it will help us to better understanding sports data analysis.

2.b Brief Overview of Literature Reviewed, Discussed and Applied

This research try to figure out a problem about how data is used in sports athletics analysis. The Specific content relative this problem includes data analysis method, model, and practical stories. Based on this core problem, here has there hypothesis in this research. When we exploring the applications of data in athletics, we can prove how these hypothesis is true or not. Because the purpose of this research paper is to help people who are interested in sports data analysis and understand sports data analysis. So we will based on this goal to do some prediction for development of sports data analysis in the future. These topics can help people who interested in sports data analysis to provide a macro and detailed reference. The research result will demonstrate value of sports data analysis. The value of research can help the people who interested in sports data analysis make better decisions about whether to conduct further research in this studying field.

Based on these goals, I found these books as my main researching literature. Thomas A. Severini's Analytic Methods In Sports (first edition), Ben Taylor's Thinking Basketball, Dr. Lorena Martin's Sports Performance Measurement and Analytics, and Wayne Winston's Analytics Stories. These literatures can support me finish this research. The book Analytic Method In Sports using mathematics and statistics to understand data from baseball, football, basketball and other sports. The main content around mathematics and statistic and probability methods and calculation in sports field. And all of data relative different sports athletics data. The book main

idea about how to use statistics and probability to analysis sports data. By reading this book, I found the big value about application of mathematics in sports data analysis. Ben Taylor's Thinking Basketball provide some thinking logical about no-number sports analysis methods. Standing on basketball views, the book gives us a lot of insights on basketball analysis from where the data can't be seen. We can get some stories from this book data analysis. This book also tell us many analysis and strategy trap in basketball data analysis. The data needs to be interpreted correctly, and the correct is made based on the interpretation result. The book Sports Performance Measure tell us many data analysis of athletes' performance measure. I will reference basketball dunk analysis to know how to measure a basketball players performance. We should know we should choose different measurement methods and measurement focuses according to different sports. We need to develop different analysis strategies according to sports. These reading material have different views of sports data analysis. They provide different views let us understanding sports measure world. But they also have some disadvantages in sports data analysis. These literatures did not make a comprehensive interpretation of sports data analysis world. For example, when we analysis a basketball game, we are not only analysis team data, but also including individual player game data, player body conditions measure and analysis, and team playing strategies. These all belong to sports data analysis world. If we put all of these thing in this research, the readers will more easily understanding sports data analysis world.

The conclusion from my literatures, including these points. 1) Sports data analysis needs mathematics as a strong support. A certain basic mathematical ability is very important for sports data analysis. Especially, as sports data analysis with ability of mathematics of statistical and probability. 2) qualitative analysis play important role in sports data analysis. Sports data analysis not only need a result of data analysis, but also need right explanations of result. So the researchers need know some basic some sports knowledge such as game rule, meaning of specific data. Because we need to transform data result into specific action and plan. 3) Data analysis not only including team and players games data analysis, but also including individual player physical measure and training performance data analysis. For example, in basketball players, physical measure including hand size, vertical jump record.

Chapter 3

3.a Study Method and Study Design

The study design (research design) including these relative content. 1)Data type required for research ,2) Research resources, 3) Participants required for research 4) Study Method (Data analysis methodologies),5) The location and timescale for conducting the data, 6) The time period required for research.

The data type required for this research including quantitative and qualitative data. Quantitative data and qualitative data including primary data and secondary data from literatures and reading materials. Based on research goal and problems, I made some virtual primary data to support some analysis conclusion. The virtual primary data can help readers better understand the progress and conclusion of problems. The virtual primary data was made by basing on real data research from literature resources. So the virtual primary data has characteristics. For example, in one of our literature, the author uses real NBA basketball player data to do comparison, I will create reasonable virtual data and un-real player name to replace author real data. I also collect some primary data from myself like my hand size measure. The some secondary data is collected from the internet and literatures. The secondary data belong to referenced data. The main data types including integer, floating-point number, and string.

The research resources except main literature and reading materials, I also reference from some internet resources for further exploring sports data analysis

world. The extra main research resources is Google Scholar. The reason for using Google Scholar is because the research requires some additional paper support detail content and argument. The all of extra resources are in the Reference page.

The participant for this research only me. I as this research author, I finished all of jobs about this research paper. My actions for this research including confirm research goals, research problems, literature research and summary, and final writing and finalization of the paper.

Study Method (Data analysis methodologies) combines qualitative research and quantitative research in this research. Qualitative research about identify developing prospects and predictions of sports data. Qualitative research methods involves collecting and analyzing non-numerical data to understanding sports data analysis. By qualitative research, I made some summary and conclusion from literatures and research resources in some sports data analysis stories and case study. Quantitative research method involve how data is be used and analyzed in sports athletics data analysis. Quantitative research method is the process of collecting and analyzing numerical data. In this research, we need to use and compare many numerical data to get conclusion and research result. In quantitative data analysis methods, we have two analysis level – Descriptive analysis and Inferential analysis. I use Descriptive analysis to find absolute numbers to summarize individual variables and find patterns. Taking some instances, Mean, Median, Mode, Percentage, Frequency and Range. I use Inferential analysis to show the relationships between multiple variables to generalize results and make predictions. Taking few examples, correlation,

regression and Analysis of variance. Correlation is describes the relationship between two variables. Regression is show or predicts the relationships between two variables. Analysis of variance is tests the extent to which two or more groups difference with each other. Specific learning and research methods including literatures reading and summary, search relative online research paper focus on sports data analysis. These research methods give me a comprehensive understanding.

The location and timescale for conducting the data. Most of data come from literatures and reading materials, a few part of data is virtual data. The extraction and application of all data determined along with the progress of the paper.

The research time period required 3 month or longer. I need to working about 20 hours every week to reading literatures and other necessary and helpful research paper to support me to finish this paper. In addition to the necessary reading, it takes a lot of time to conceive the structure and content of the paper. I takes about 3-4 weeks to write and finalize the research paper.

3. b Explanation of Sample to Be Used in the Study

In this research, I mainly quoted two main data samples. That is Kevin Durant 2011- 2012 season basketball data and NBA Sun's 2005 Stoudemire and Hunter shooting data. These data all belong to individual player basketball performance data. These data as samples can represent players one moment or one season specific performance.

According Kevin Durant 2011 -2012 season basketball statistics data, I used it as example and real case to analysis mean, stander deviation and margin of error in sports data sample. The sample includes 6 parts, they are statistics, rebounds, assists, turnovers, fouls, points. The data samples was cited from literature. It is reliable and validity.

Another data sample about 2005 NBA Sun's two players Stoudemire and Hunter. The samples comparing shooting percentage at the different area between them. Stoudemire has better shooting percentage wherever closer basket or in the middle range. Hunter has the same data in area where closer basket, but mid-range shot is lower percentage than Stoudemire. These samples was used to explain how to correct understanding basketball analysis theory.

3.c Explanation of Measurements, Definitions, Indexes, etc. and Reliability and Validity of Study Method and Study Design

Basically, based on how the data and samples are obtained, the research data measurements and indexes are reliability. Data has the characteristics of long-term validity and not time-sensitive. The data represents the solid characteristics of sports data, which is reasonable and measurable. The data can represent the performance characteristics of a type of athletes, but does not represent the performance characteristics of all athletes. The part of definitions are reliability. Because as the sports data analysis technology developing, probably some definition would be changed in the future. The virtual data and real data are reliability and reasonable. The real data from true sports cases and athletics performance. The virtual data mimic real data, so virtual data is reliability. The real data on such as sports players performance, the measurement is reliability and validity.

The some definitions and relative research conclusions are not reliability. Because, some research conclusions need time-consuming test. Because time limited, some data analysis results probably have deeper exploring and finding. Some definitions come from some phenomena and theories. But not one hundred percentages assume all of theories and phenomena are right. Probably, one day, some hypothesis has new finds and original theory structure would be changed by some new finding.

Study measure and study design is reliability, The reasons same with data reliability. The data from sports case history, history won't be changed. So study measure is reliability and validity. About study design, probably exist some little or more problems, so reliability and validity are determined by time and readers. The research techniques are based on reading a large number of literature and summarizing the existing theories. Analytical techniques combine personal understanding of the literature and use induction to arrive at a new point of view. From the analysis of the research results and research goals, this study has basically completed my expected goals, so it is concluded that research design and research methods are reliability and validity.

Chapter 4

4.a Results of the Method of Study and Any Unplanned or Unexpected Situations that Occurred

The results of the method of study are comprehensive. Based on method of study, the research figure out all pre-set problems and fully demonstrates hypotheses. From the research result, we find multiple statistics and probability method be used in sport data analysis processes. Sports data analysis need to using mathematics and statistics to understand data from different sports program. By the methods of study, the sports data analysis have a detailed steps includes describing and summarizing the sports data, probability and statistical methods to analysis sports data, and using correlation to detect statistical relationships. Other methods including modeling relationships using linear regression and build regression models with different predictor variables.

By analyzing basketball, we also figure out different some sports analysis logical. For example, Braess's paradox. If a high performance players always increase number of shots, this behavior is bad for whole team scores.

Sports performance measurements belong to part of sports data analysis. Here has five important physical variables includes stamina, speed, strength, skill, and spirit. Based on different sports program, these variables play different roles. Except these necessary factors, sports player psychological still need to measures.

By analytics stories in sports, we also get some other conclusion about sports data analysis reliability and application. These stories give us a professional view to see sports data importance. From these stories in real sports case, we make a Conclusion about development prospects of sports data analysis.

About Unexpected and unplanned situations, one of situations is research time is limited. Underestimate the time it takes to write the research paper. The research paper structure is complicated and some content need time to learn how to write. If the time is sufficient, the conclusions of the study may be more reliable and valid. Another is some other factors what influence sports player performance are difficult to measure and analysis. Sports data not just games statistics and athletes physical measures, psychological status can be measured, but how it is hard to measure effectiveness. In study method, I found data sample collection by multiple methods. Normally, we got sports data by sports media, public datasets for history sports recorded, or watch games, collect data by yourself by measures and recoding. But some methods need hardware support, the funding is limited. To solve this problem, I compensate for this problem by reading the literature and try to build comprehensive sports data analysis world. About literature, some contents are very useful for this research, but part of content not necessary to reading. About methods of study, underrate the importance of practice. For some questions, the Excel, R and Python should involve in method of study.

4. b Explanation of the Hypothesis

Research hypothesis is a specific, clear, and testable proposition or predictive statement about possible outcome of a scientific research study based on a particular property of a population, such as presumed differences between groups on a particular variable or relationships between variables. According the research hypothesis concept and research goals, I provide two hypothesis about sports data analysis. The hypotheses include Data can replace coaches to guide games and sports data analysis meaningful for predicting the outcome of the game.

These two hypotheses research method is qualitative data analysis. By reading literatures and other research materials, I combining my personal experiences and research resources to make the conclusion for these hypotheses.

Hypotheses are appropriate knowledge expansion in this research paper. By discussion of hypotheses, we can deeply understanding the problem we are researching. For the first hypothesis data can replace coaches to guide games, this hypothesis stand on different view to talk about what is data role in sports data analysis, what is relationship between technology and human, what is data analysis biggest potential. All of these questions we all have related discussion in this research. The second hypothesis data is meaningful for predicting the outcome of the game. We discussion sports data analysis reliable in the research. Because the sports data analysis include probability, sports betting is huge part in sports data analysis for predict outcomes of sports games.

Chapter 5

5. a Full Discussion of Findings (Results) and Research Analysis of Finding

The sports data analysis by using of mathematical methods, together with the vast amount of data now available, to analyze performances, recognize trends and patterns, and predict results. To provide insights on athletes and teams, we need to developing different mathematics analytic methods on such as statistics methods and probability. Statistics concepts are core in sports data analysis. The goal is to taking advantage of methods to extract useful information form the data.

Beginning of data analysis, thinking about subjects and variables. Subject is an topic which data do you need or which data are collected. In sports data analysis, usually, the subjects are athletes, sometimes the subjects are games, seasons, teams, or even coaches. Defining subjects is a key point that know what we want to do by analyzing data firstly. After defining a subject, the next step is considering about variable. Variable is a characteristic of a subject that can be measured. For example, if we want to measure a NBA basketball player games performance n 19 - 20 regular season, we need to know the player's played how many games, how many minutes, played with which teams, scores, assists, rebound, steal, and block are all examples of variables. Variable has an important property, called it measurement scale. Usually, the measurement scale is a set of numbers. For example, if our subjects are NBA all Centre in 2019-2020 regular season, the variable

is the number of how many blocks in this season, then the measure scale is the set of integers 0,1,2,3….,30,31,32… Variables whose metrics consist of numbers are considered quantitative. We need to care about not all variables are numerical. For example, if our subjects are MLB players in 2019 season and variable is the players bats, and then the variable values is left(L),right(R), and switch hitter(S). Some variable such like these is called qualitative or categorical.

Analytic method means using data to extract conclusions and make decision. But the data not always available and clear. So we need to filter out the noise and clear data to see deeper relationships in the data. Sports data analysis belong to observation study, the study not be control by data analytics, because we can't generate data from our ideal ways. For example, we can't control which two players or teams participate in a one game under a given situation. We have to accept all data that randomly generated from the games or players. For different data properties, we should use appropriate type of analysis.

In usually, the first step is some type of summarization in analyzing sets of data. The simple tool in summarization is frequency distributions. For example, we can build a frequency distribution table to know a team performance (lose and win) in the whole season.

Table 1 AAA high school wins and losses in 2019

W	W	W	L	W	W	L	L	W	L
L	L	L	W	L	W	W	L	W	L
W	W	L	L	L	L	W	L	W	L
W	L	W	W	W	L	L	L	L	W

Table 2 AAA high school Win-Loss Record in 2019

OUTCOME	NUMBER	PERCENTAGE
WIN	19	47.5%
LOSS	21	52.5%
TOTAL	40	100%

When we using frequency distributions, we need to choose appropriate classes. At the same time, we need to let classes range in the same length. For example ,see Table 3.

Table 3 XXX football quarterback passing yards in games started, 2015-2019 season

CLASS	COUNT	PERCENTAGE
LESS THAN 100	20	32.8%
100 - 199	35	57.4%
200 - 299	34	68.9%
300 OR MORE	3	5%
TOTAL	61	

Of course, we have many other ways to present frequency distributions. For example, Histogram. Except using frequency distributions to know a team or personal player performance in an entire view, mean and median in many cases is more useful and most common for quantitative data. The median and mean are important for summarize a set of data in different situations depend on goals of analysis. Standard deviation is another way to know and analysis variation in sports analysis. The standard deviation can reflect the degree of dispersion of a data. Normally, smaller SD means a more stable performance, and vice versa. Another analysis variation way is IQR (interquartile range). IQR is the upper quartile minus the lower quartile. In sports analysis, IQR has two advantages, one of advantages it more direct interpretation that is often useful in understanding the variability in a variable. Another advantages is IQR not sensitive to extreme values than the mean.

By analysis mean, median, SD and IQR in a dataset of variations, we can have good view to familiar our data conditions for entire analysis.

Probability theory is the branch of mathematics that deals with random outcomes. So probability play important role in sports data analysis. Because we can't control random element in sports. Understanding probability is benefit to you to draw the good conclusion from the sports data. Probability is means any process that produces random results. For example, a basketball player plays free throw. Nobody know whether he can get another free throw score. Some examples like this is lots of in sports field. In probability includes more or less events. For instance, if X is a possible event, it is usually demoted the probability of X by P(X). P(X) = 0.25 is mean event X will have 25% probability happen. For example, if we have event A and event B, and A, B can't occur simultaneously, A represent a player hit a single, B represent the player hit a double in baseball game, and then let us hypothesis P(A) = 0.3 and P(B) = 0.04, the probability that hitting wither a single or a double is P(A or B) = P(A) + P(B) = 0.3 + 0.04 = 0.34. This is a most common way to consider two events happen probability. Taking one more example, a hypothesis event C represent a baseball player hit single in 0.20, denote P(C) = 0.2, then probability that he does not hit a single is P(not single) = 1 – P(C) = 1 – 0.2 = 0.8 However, in sports data analysis world, we not only care about events, also care about numbers. When we have huge number of numbers data, probability can help us interpreted as long run relative frequencies in a large sequence of experiments. For example, from

history records in win-loss results between football team AAA and team BBB, team BBB has 30% probability will win the next game. But sometimes, we will consider other events together to predict something. For example, P(B) represent team BBB win football game probability is 30%, denote P(B) = 30%, quarterback is critical player in football game, QB throws 5 interceptions, the team BBB win about 3% of all the time. So P(team BBB win | QB throws 5 interceptions) = 0.03. This is conditional probability. Conditional probability is very useful because it allow us to incorporate other assumptions or events into probability. Anyway, when we do sports data analysis , probability can help us better to understanding factors in sports events and predict some event happened probability. This is a very useful analysis logical in sports analysis. But how to use it should be combine different situations and sports programs. Depending on this point, a good familiar for a sports program is benefit you to do draw more meaningful data analysis conclusion. For example, in soccer game P(home team wins) and P(home team score first) always put together to analysis and predict game results such as P(home team wins | home team scores first)

Another necessary mathematics method is Statistics in Sports data analysis. The estimation is part of statistical. Estimation means using data to determine the properties of the underlying probability distribution, it also called statistical estimation. For example in football game, we want to know a Running back performance in every carry. Let us make a hypotheses, the running back player name is R. The R every carry data denote R1, R2, R3…, All of these data sample mean of R is

5.59 yards per carry, so we estimate the future carry data about 5.59 Similarly, the sample median of the R values is 3 yards per carry, we also estimate of the median of R. Normally, we like using the margin of error to quantify the variation in sports statistics. For example, a running back R per carry value average (mean) is 5.38, the margin of error is 1.16 So the interval gives a range of values such as 5.38 -1.16 to 5.38+1.16. This range can tell us a player true performance and summarize a player performance. If we have lots of data and the game will repeat, we can use this value to predict a player performance in the future. The uncertainty under this prediction is also be described by the margin of error. About margin of error calculation, there have two ways. One of ways is to use statistical formulas.

$$2\frac{S}{\sqrt{n}}$$

S represent sample standard deviation of a player game-by-game points scored and n denotes the number of data values used in the sample average. For example, we using NBA super star Kevin Durant's 2011 -2012 season data to see margin of error

Table 4 Kevin Durant's 2011 -2012 season data.

statistics	mean	SD	MARGIN OF ERROR
Rebounds	7.98	3.03	0.74
Assists	3.50	1.95	0.48

Turnovers	3.76	1.70	0.42
Fouls	2.02	1.36	0.34
Points	28.03	6.88	1.69

As we all know, suppose we know the standard deviation values of each data, for convenient calculation. For example, depending on our formula, we get margin of error values. For Points section, Durant points range between 29.72 and 26.34

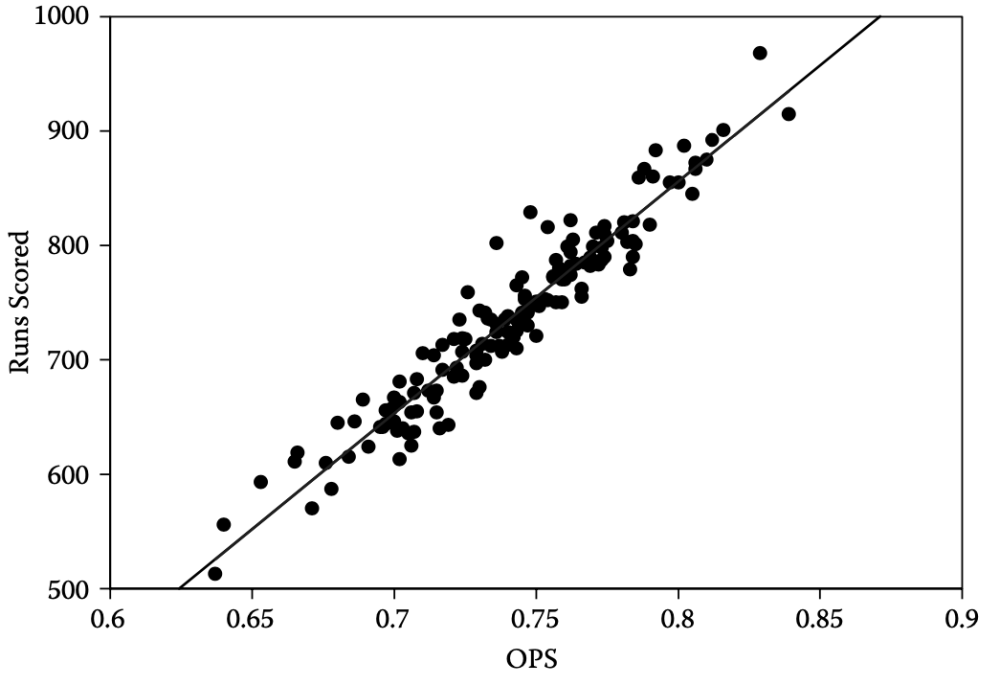
In some sports ,we probably need to using proportion in some probability events. Under this conditions, we can use this formula to calculate. P(A) = p. p is the proportion of experiments in which event A occurs. The n is same meaning with last formula.

$$2\ \frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

Another way is compute the margin of error using computer simulation to obtain these hypothetical repetitions. You can find this way easy on internet. This paper not mathematics paper, so we do not talk too much content about this approach in these.

We need to use some models to detect strength of the relationship between variables and presented. These approaches can help us reduce the properties of such a relationship to a single number that is useful as. A simple summary of the relationship between variables. The most common approach is linear regression. We

also called it as linear relationships. For example, we measured two variables Y and X which we are interested and we want to know what relationship between these variables. We let Y represent baseball runs scored and X represent on-base plus slugging(OPS). By linear regression analysis, we know these variables have a strong relationships. If not, the points in graphic are very mess up .



The simple linear regression model relates a response variable Y to a predictor variable X. Sometimes, if we meet multiple factors influence variable, we called multiple regression analysis.

$$\hat{Y} = a + b_1 X_1 + \cdots + b_p X_p$$

Based on our current discussion, we see that sports data analysis is not a completely independent discipline. Sports data analysis is a research field based on mathematical foundation and flexible use of statistics knowledge. Many methods and theories of statistics are the basis of sports data analysis. Therefore, if you want to study sports data analysis in depth, a solid mathematical foundation and rich statistical knowledge are indispensable. From the perspective of mathematical statistics, when analyzing sports data, we go through the following steps, the first step is find data from relivable and available sources. The Second step is clear data, clear noise in data and choose necessary data for analysis. The third step is thinking about subjects and choose variables for goals of subjects. The forth step is using mathematics statistical methods to analysis data and draw appropriate conclusions depending on subjects goals and variables. When we analysis data with statistics methods, according to different purposes, we should choose the appropriate methods such as mean, median, standard deviation, linear regression. Basically, statistical knowledge is key methods to analysis sports data in quantitative analysis.

Sports data analysis not only quantitative analysis like statistics methods, but also need some qualitative analysis by summarizing specific sports event. In this research, we focus on basketball program.

Wilt Chamberlain attempted more than 25 shots per game in 1966. He is the best player and most efficient scorer in the team. Normally, we would like asking the team most efficient scorer to shoot constantly. The reason is very easy, he is the best, he got the most scores, he should increase numbers of shot and then team win.

But in fact, this idea not work in team sports like basketball. The Braess's paradox can explain why Chamberlain shouldn't constantly shot. Braess's paradox is a phenomenon that has been observed in cities where a major highway closed. When a city without an obviously superior highway option, drivers naturally distribute themselves more evenly across other different roadways. In basketball, distributing shots among teammates is like dispersing cars across different roadways. The competitors' defenses same with other cars, these cars influence the optimal route home, basketball defenses influence the optimal scoring path. The opponents must be not allowed your best player constantly get scores, they always thinking and switch different defensive strategy to influence your best player. Meeting these defensive, few players can keep high-percentage scoring performance instantly. So under this conditions, the best player try more shots, the positive affection is less for team wining game. In team sports, we should pay attention to team's overall efficiency. Chamberlain should shoot less and increase teammates scores. This impact is called as "Global Impact". In other words, global impact is the effect of a player's actions on his teammates and himself. Braess's paradox applies to basketball because the winner is determined by the team total scores other than individual's scores.

In basketball data analysis, we have a theory about redistributed and not replaced. In 2005, Phoenix Sun has famous player Stoudemire, he scored 36 points every 100 possessions he played. But when he backup and not play, Phoenix was not

lost 36 points per 100 possessions. His backup, Steven Hunter, scored 17 points every 100 possessions, about 19 less scores than Stoudemire.

Table 5 Stoudemire and Hunter 2005 shooting data

2005 SHOOTING	AT THE RIM	MID-RANGE
STOUDEMIRE	69%	44%
HUNTER	69%	28%

From the data analysis, we see Hunter almost no mid-range capability but he almost same scoring ability with Stoudemire at the rim. When Stoudemire not on court, Phoenix not find a way to replace his scores, they find a different ways to redistributed the scoring options. So when Hunter on the court, Phoenix find another way such as he closer to the basket, and teammates pass on him. The new distribution of scoring attempts can help team get better effective probably. Because nobody can replace Stoudemire get scores effective.

When we analysis a players scoring abilities, we can’ t only see player’s data such as how many average points per game and how many points per scoring attempt. These data can represent something but not all. At least in basketball, we should think more than that.

Table 6 Player A and Player B virtual data

Player	Average Points per game	Points per scoring attemp

A	11	1.00
B	20	1.10

Normally, after we analyzing this table, player B is better scorer than A, But in fact, he is not.

Table 7 Player A and Player B virtual data

Player	shot	Idle Pass	Create jumper	Create layup	Team points
A	11	39	10	5	67
B	18	39	5	1	65

From this table, you will know player A is a creator and he help team get more scores than B. Player B indeed is a good scorer, but under global impact theory, Player A can lead team get more wins than B. Because A creates more teammates scoring opportunities. We can't interpret data from the surface, how to use data and get a useful conclusion to help team or players, that we need to do by sports data analysis. Thus occupying defenders to create open shots, even for less skilled teammates, is really valuable; an open shot for a role player is often more efficient than a covered one for a star. From this theory, when we analyzing sports data, we need to combine data meaning and sports program characteristics for a good conclusion.

Some conclusion even violation of common sense in sports data analysis. Like I just talk about cases. Player A has less shoots, but the team overall increase five

points for efficiency. This improvement was not accomplished with different players, but with a different distribution of scoring attempts. This is also Braess's paradox application in basketball. But the same conclusion probably not work in other sports programs. So understanding sports program character is very important in sports data analysis. If you only know how to analysis data, this is not enough to be a sports data analyst.

Basketball is a per-possession game. By rule, teams alternate possessions – one team's offensive possession ends in one of four ways: 1, made shot 2, a turnover 3,defensive rebound 4, end of quarter. Because of these alternating rules, basketball success is completely determined by per-possession efficiency. When we do basketball data analysis, we should consider about causal factors. Ask why player get high points? who give player assists? Who are opponents? For example, if Nash not on the court, Stoudemire's easy layup would be disappeared. From this point, we need to pay attention to scoring causal for every possessions. Finding causal reasons will help us the truth about scoring. Individual scoring is a one of factor influence team win, but the purpose of basketball is for a team to score more points than its opponent. So finding scoring causal is more important than know who is higher scorer in a team. How do we get this conclusion? We combined data and basketball theory to come up with this piece of conclusion.

When we analyze sports data, we should avoid bias and consider Anchoring fluences. The human brain requires a starting point. If we very care about one starting point, we will easy ignore other factors or event. For example, in sports data

analysis, we would like thinking scoring capability is the most important metric for a basketball player. For most of basketball fans, this view is feasible. Because scoring became the default measure of a player's contribution to wining and fans feel like easy understanding it. But for sports data analyst, this view will let us generate bias in data analysis. As sports data analyst, we should comprehensive analysis sports data and reject preconceived ideas and opinions. Anchoring is a phenomenon, the order in which information is presented matters. Meeting similar information in different order, we will produce different judgment linked to the original information. How to avoid or less anchoring phenomenon? Adjusting information and adequately new information. For example, in basketball, we should pay attention to efficiency rather than volume of scoring. At the same time, we should pay attention to global impact, an individual player how influence other teammates, rather than consider individual who score.

When we do data analysis, we should care about data details. For example, in basketball, it as a high-volume sport, the small differences compound to produce large discrepancies. When we calculated and analysis every 100 possessions data in basketball, the small difference in percentile multiplied by one hundred will also become a whole number.

The number of samples collected should be determined according to the characteristics and attributes of the sports. For example, because of basketball is a high-variance nature, it should takes a larger sample size, a large number of games - to be confident that the numbers are actually reflective of the overall performance.

In sports data analysis, the data variance and sample size usually show up together. The rule about sports data sample size is the greater the variance, the larger the sample needed to make accurate conclusions. Because when we to confirm a team good or bad, not only analysis one games or one series, we need all data! The richness of data samples is conducive to the accuracy and reliability of the final analysis results. At the same time, we need to pay attention to sample-size insensitivity. Sample-size insensitivity meant a tendency to consider the given sample as sufficient for reaching a conclusion. Because of this we shouldn't ignore importance of sample size in sports data analysis.

Sports data analysis also including physical assessment. Understanding body measurements can help improve the athletics performances and also help improve training efficiency. As sports performance analyst, you should know some necessary measurement methods for your physical performance variables. The physical variables includes body composition, muscular strength, power and endurance, flexibility, balance, anaerobic power, aerobic power, reaction time, agility, and level of sport-specific skill. The some fundamentals of fitness measures include muscular strength, muscular endurance, flexibility, cardiorespiratory fitness, and body composition, muscular power, coordination, balance, and anaerobic power. All of these factors are measured as accurately as possible. In physical measurement, we need to use some necessary formula. For example, Body composition includes muscle, fat, bone, and other substances in the athlete's body. Normally, we use BMI

to calculate and individual athlete fat content based on the relationship between weight and height. BMI is calculated as the weight in kilograms(kg) divided by the height in meters squared.

$$\text{Formula for BMI:} \frac{\text{weight in kilograms}}{(\text{height in meters})^2}$$

According to the measurement of body composition, the corresponding measurement method and evaluation methods should be adopted. In this process, you need to refer some relevant forms and tables to draw conclusion. For example, this table built for assess muscle strength, muscular power and muscular endurance.

Table 8

	Muscular Strength	Muscular Power	Muscular Endurance
Leg Press			
Have the athlete place the feet shoulder width apart on the footplate. Have him or her form a 90-degree angle between the hips and knees. Finally instruct the athlete to keep the gluteal muscles on the seat throughout the entire range of motion of this exercise. Measures quadriceps, gluteus, soleus, and gastrocnemius strength, power, and endurance. This is more of a muscle group assessment.	Use 1RM or submaximal loads; 4RM, 6RM, or 8RM	Perform 3 repetitions as explosive as possible (during contraction phase of exercise) between 45-65% of 1RM	Perform as many repetitions as possible until fatigue or up until 2 minutes using 20-35% of 1RM
Leg Extension			
Have the athlete sit with the seat adjusted so that the knees are in line with the machine's axis of rotation. Then instruct the athlete to perform the exercise by kicking out until the legs are almost fully extended and then return back to starting position. Another measure of quadriceps strength, power, and endurance. This is a more isolated muscle assessment than leg press.	Use 1RM or submaximal loads; 4RM, 6RM, or 8RM	Perform 3 repetitions as explosive as possible (during contraction phase of exercise) between 45-65% of 1RM	Perform as many repetitions as possible until fatigue or up until 2 minutes using 20-35% of 1RM
Leg Curl			
Have the athlete lie prone on the pad and align the knees with the pivot point of the machine. Next, the athlete should be instructed to begin the exercise by bringing the pad toward the buttocks, holding the movement, and then lowering the weight slowly without raising their body off the pad. This is a great assessment of the hamstring muscles and muscles in the gluteus region.	Use 1RM or submaximal loads; 4RM, 6RM, or 8RM	Perform 3 repetitions as explosive as possible (during contraction phase of exercise) between 45-65% of 1RM	Perform as many repetitions as possible until fatigue or up until 2 minutes using 20-35% of 1RM
Chest Press			
Have the athlete perform this exercise by pushing out on the handles until the arms are nearly straight and then return to starting position. Adjust the handles so that they are at chest level during the initial part of the movement. This is a great assessment for upper body muscle group including pectoralis major and minor for strength, power, and endurance.	Use 1RM or submaximal loads; 4RM, 6RM, or 8RM	Perform 3 repetitions as explosive as possible (during contraction phase of exercise) between 45-65% of 1RM	Perform as many repetitions as possible until fatigue or up until 2 minutes using 20-35% of 1RM

This is part of form about body muscle group testing for athletes. This form give us a instruction about how to test and assess muscle for athletes. Some knowledge about

physical measure and analysis should refer some special tables. So in physical data analysis, the table for reference is important. For example, when you need to make conclusion about motion of single-joint form your data, you should rely on range of motions of select single-joint movements in degrees. This knowledge is solid and fixed.

Table 9

Movement	Degrees
Shoulder Girdle Movement	
Flexion	90-120
Extension	20-60
Abduction	80-100
Horizontal abduction	30-45
Horizontal adduction	90-135
Medial rotation	70-90
Lateral rotation	70-90
Elbow Movement	
Flexion	135-160
Supination	75-90
Pronation	75-90
Trunk Movement	
Flexion	120-150
Extension	20-45
Lateral flexion	10-30
Rotation	20-40

Physical analysis belong to sports data analysis, and it needs anatomy and physiology knowledge. Basic anatomy and physiology is fundamental to obtaining a more comprehensive knowledge of what these physical data means to be an athlete. For example, the composition of the human body, distribution and function of muscles, and types and functions of bones. The analysts need to collect and measure by themselves for physical analysis. During this process, analysts need to learn how to collect data by some equipment and machine. Sports data analysis need technology hardware support for collect data. In order to better service athlete and team,

physical analysts need understand the requirements of specific sports for various indicators of physical measurement. For example, basketball player need sports-specific skills includes precision, accuracy, and consistency of shooting free throws, three pointers, and passing the ball precisely to teammates, adaptations for other player positions are recommended. But the tennis is different with basketball. Tennis requires players precision, accuracy, and consistency of ball placement should be measured for each of the major tennis stokes: serve, forehand, backhand, volley, slice. Because these sports characteristics of tennis, the field of tennis has some new technology to measure these sports and body moving data. For example, installed sensors within the grip handle of the racquet to quantify measures specific to tennis players. The sensors records the frequency of strokes, type of spin used by a player, and racquet speed among other measures.

For basketball, the athletic tests include a standing vertical jump, maximum vertical jump, bench press, three-quarter-court sprint time, lane agility time, and modified event time. The physical measures include height without shoes, wingspan, weight, standing reach, body fat, hand length, and hand width. These measurements have a strong pertinence to the characteristics of basketball. In the process of data analysis, you can use the method of comparative analysis to compare data with similar basis conditions. For example, the same body heigh players, comparing them hand size.

In addition to physical analysis, these is also psychological analysis. In sports data analysis, as data analysts, we should know our data. Because data is

fundamental to understanding the factors that can play a role in the athlete's or team's performance. Psychological also play important role in the analysis. Psychological factors will cause variance in physical performances. Sports data analysts also need to measure and quantify these intangible variables by some models. In sports the key psychological factor that influence player's performance including motivation, confidence, anxiety, depression, aggressiveness, self-esteem, self-efficacy, and concentration.

Regarding the measurement methods, some measurement models that have been widely recognized in the academic world, especially in sports psychology. Based on your measurement purposes, there will be corresponding measurement models have been repeatedly used and proven. For example, measure motivation. As we all know, the strong and positive motivation that affects sports performance, whatever it comes from coaches or the athletes. The motivation has two main subcategories, one of subcategories is intrinsic motivation and another is extrinsic motivation. The Intrinsic means "I want to do" for satisfy some satisfaction. The extrinsic motivation depends on external rewards. The sports motivation scale developed in France in 1995. The original scale measures both intrinsic and extrinsic motivation but the effective of original scale was questioned, and then developed new scale SMS-6(Mallett et al. 2007). Now, in order to determine specific situation motivation measures, the scale is SIMS (Situational Motivation Scale)(Guay, Vallerand, and Blanchard 2000). From this example, we know sports psychology measure by using some specific scale that have existed. We need to note that these

measurements are not static. With the progress and development of academic research, sports data analytes must learn to adopt new measurement scales for different situations.

5. b Full Discussion of Hypothesis and of Findings

The first hypothesis of this research is data can replaces coaches to guide the game. The final conclusion is that data cannot replace coaching in person. This final conclusion of the hypothesis is the opposite of the initial judge. As we all know, with the developing of technology, the machine learning and artificial intelligence paly more and more important roles in different field. The same is true for machine learning and artificial intelligence. But data can't take advantage of machine learning and artificial intelligences to coaching human to play game. At least data can't guide team competitive sports. Data can't play coaching role to tell players how to play game and replace human players in live games.

Competitive sports like basketball, soccer, they all have some specific features including uncertainty of the game, accidental events during the game, the nature of the win based on the total score of the team. Because of these factor influence games, the coaches job nature is complexity. Some uncertainty events are an important part of the game, no one wants to watch a game with a pre-set result. The machine learning can learn lots of game strategies for one sport, but human players performance can't control by anyone for example, injury and the arrangement of opponent's tactics. Machine learning can learn these strategies by supervised learning. But teams competition not only including playing strategies , but also including player mental recognition and psychology analysis. The coach should to comprehensively know players conditions and status. From this point, data can't do that. For example, the coach's not to put five highest-scoring players on the field, but

to put the five most suitable players on the field. The coach is going to motivate the players on spiritual. Yes, machine learning also can learn these motivation languages, but whether the players accept is a big problem. The complexity of the game and the complexity of the coaching work determine that the data cannot fully replace the coaching work in person.

From the machine learning technology view, the same question like whether the autopilot can achieve level-5. Level-5 means full automation. The vehicle performance all driving tasks under all conditions. Zero human attention or interaction is required. Complicated road conditions are similar to complicated games. Machine learning and artificial intelligence can't analysis and handle all of these uncertainty and accident conditions. Autopilot deep learning need pre-learning process. But meeting some conditions they haven't experienced before like snow or ice covered- road. As human being we can use intuitive physics to handle these conditions and make an appropriate decisions. And as human, we can rely on our knowledge of how the world works to make rational decisions when we deal with new situations. We also can causality and can determine which events cause others and understand the rational actors for our next move. For example, when you driving and see a rhinoceros on the street. Under this conditions, you will try to find a new road. Even though the rhinoceros walks slowly and it won't touch your car. But for the same conditions, deep learning algorithms don't have such capabilities, therefore they need to be pre-trained for every possible situation they encounter.

Data support machine learning, and machine learning can help robot or machine more intelligence. But data requires specific application environment to make the appropriate decision through analysis. We can use data to analyze games and players, but it is not feasible to use data to replace coaches and coaching. From the sports view, using a computer that analyzes data to replace coaches does not meet the recreational nature of sports. When we watching games, we would like to watching coaches body language ,facial expression and some funning words. If a computer or robot that uses data knows about the game, it may be difficult to achieve or the entertainment nature will be greatly reduced.

The second hypothesis is that data analysis is meaningful for predicting the outcome of the games. The hypothesis is true. First of all, the meaningful means that it works, can achieve certain expectations, and has a certain affect. Sports Data analysis play an important role in sports betting. As we all know, the outcome of sports game is probability and unpredictable. We don’t discuss issues related to illegal competitions and violations of the spirit of sportsmanship. If we talk about sports data analysis meaningful for predicting game results, we have to talk about statistics. Sports data analysis need statistics methods to draw conclusion. In other words, statistics is benefit to betting on sports. By statistics, we need to identify the factors that have a strong correlation to wining games. Sometimes, these factors aren’t immediately apparent to the betting public world. The analysis process is long or hard, but the reward will be worth it.

When we talk about statistics in sports betting, we should know “significance” does not mean “important”. We need significance data from our dataset. The first step we need what factor significance influence game win or loss. For example, if we want to know a NFL team wins or loss, the “completion percentage” is significance data. And then we need to find a dataset includes these data. “completion percentage” is independent variable, the game win or loss is depending variable. The more statistically scientifically important a variable is, the more likely you are believe that is related to winning. In sports data analysis, here always have many variables at play at once, so statistics method multiple regression is most commonly used for sports betting. For example, for predict a basketball game result, we need multiple regression. The guest team won the last two games by two points in home, the home team win 93% of games in which they score 105 or more, the home team have won 92% of their games at home, these information all belong to multiple regression analysis. In sports betting analysis, we always use historic data to predict future outcomes of games.

Another statistics way is logical regression. Logical regression is a method for get result from one or more independent variables. For example, a basketball team 3-points percentages, the total number of offensive rebound, and the total number of assists have an influences on the probability winning? When we analysis variables, we need to pay to attention to correlation and causation. For example, variable A and variable B. You can say variable A and variable have a correlation, but they are not necessary have causation. Regression analysis can help us to find variables that

correlation such as home field advantage and winning percentage, but we can't say winning percentage is caused by playing at home.

Except statistics, the sports data analysis also need probability knowledge for sports betting. The main methods include the Bayesian Network, Poisson distribution, and binomial distribution. All of these methods can help us to predict outcomes of games. These methods also determine true in the data analysis meaningful for predicting the outcome of the games.

5, c Post Analysis and Implications of Hypothesis and of Findings

According to research, I found the most of data analysis practical methods by using Excel and R. Statistics are core in sports data analysis. Excel and R have strong usability for data analysis. With developing of data analysis technology, I believe more and more new technology will be used in sports data analysis. Based on correct understanding of sports data analysis, in the future, the sports data analysts should be required more effective working for data. The data technologies comprehensive using is necessary. The data extract, data mining, data processing and data visualization, and data storytelling. In this process, Tableau, Python, Machine learning will play more and more important role in sports data analysis. As the data collection, the huge number of data need big data technology to processing. The development of Machine learning will have a direct impact on the development of sports data analysis. The greater the amount of data, the greater the advantages of machine learning. Sports data analysts need to adapt to the development of new technologies. The theories of sports data analysis will also change as the rules of game change. Continues learning ability is also necessary.

We can’t ignore hardware power. Like I said before, sports data analysis require high effective. In other words, using less time, and get more analysis results. Combining Hard wares with related software can help us achieve this goal. Recently, this technology has been used in soccer. In some countries, soccer player wear special clothes and put a sensor in clothes to detect the athlete’s heart rate and

sports performance. The assistance coaches only using iPad and reading athlete data immediately. High-tech equipment will definitely have an impact on the sports data analysis. In the future, the essential for the ability to analyze and interpret sports data is necessary. The team not only need people who master of data analysis or statistics, but also need a people who know sport nature. Therefore, the accumulation of specialization and experiences is a challenge for sports data analysis. Processing data is not the ultimate goal of analysis. Only constructive opinions can really help the team and athletes.

Chapter 6

6.a Summary of Academic Study

The academic study focus on sports data analysis application methods and processes. Explaining what is sports data analysis from multiple angles through quantitative analysis and qualitative analysis. The study involves mathematical statistics, probability theory, and other summary of sports data analysis. The study process always contains some details and lot of examples. Through the collation of sports data analysis concepts, some necessary conditions and skills for sports data analysis are summarized.

The academic study goal is the purpose of academic research is to explore the what is sports data analysis. The academic study also emphasized the characteristics of sports data analysis and the differences from other industry data analysis in the research process. For someone who want to learn sports data analysis or plan to apply data analysis in sports athletics, this academic research can help them understand sports data analysis and the necessary skills, so as to provide references to decide whether to engage in this industry. People who are interested in sports data analysis, such as fans, can learn from this academic study how the sports data analysis processes are made.

In this academic study, two questions were discussed and two related hypotheses were answered. The two questions are how data is used in sports athletics analysis and what are the development prospects and predictions of sports

data. The two hypotheses are data analysis is meaningful for predicting the outcome of the games and data can replaces coaches to guide the game. Through these two questions and hypotheses, reader can fully understand sports data analysis from a macro and micro perspective. The judgment of the hypothetical result is a personal summary based on literature and data analysis. Support the hypothetical results with sufficient arguments and examples.

The findings of academic research include the following. (1) Sport data analysis is not a completely independence discipline. Sports data analysis need strong and solid knowledge about mathematics statistics and probability. Analysts need to understanding statistics and probability theories and methods. According to different analysis goals, using different statistics methods. (2) Sports data analysis need know characteristics and nature of specific sports program. Sports data analysis need to know the significance of various variables for data analysis. (3) The results of sports data analysis need to be derived from multiple aspects, requiring a combination of qualitative and quantitative analysis. Avoiding bias and anchoring influences. (4) Sports data analysis involves the analysis of player' personal game data, team performance data and the players' individual physical conditions. The physical conditions include physical factor measure and psychology measure. (5) The sports data analysis processes including finding data from reliable and available sources, cleaning data, thinking subjects and variables, using statistics to get analysis result, and propose conclusion. (6) Psychology analysis and measure need hard ware support and reference related tables.

6.b Reference to Literature Review

The academic study focus on sports data analysis application methods and processes. Explaining what is sports data analysis from multiple angles through quantitative analysis and qualitative analysis. The study involves mathematical statistics, probability theory, and other summary of sports data analysis. The study process always contains some details and lot of examples. Through the collation of sports data analysis concepts, some necessary conditions and skills for sports data analysis are summarized. The research try to find answers about how data is used in sports athletics analysis and what are the development prospects and predictions of sports data. The research also provides argument and discussion for the hypothesis.

The academic study goal is the purpose of academic research is to explore the what is sports data analysis. The academic study also emphasized the characteristics of sports data analysis and the differences from other industry data analysis in the research process. For someone who want to learn sports data analysis or plan to apply data analysis in sports athletics, this academic research can help them understand sports data analysis and the necessary skills, so as to provide references to decide whether to engage in this industry. People who are interested in sports data analysis, such as fans, can learn from this academic study how the sports data analysis processes are made.

The goals of literature review is to gain an understanding of the existing research and debates relevant to the research topic, and to present that knowledge in the

form of a written report. The literature strongly support the research in sports data analysis. Literature review can prove the rationality of the research and provide sufficient theoretical basis to support the research problems.

Based on these goals, I found these books as my main researching literature. Thomas A. Severini's Analytic Methods In Sports (first edition), Ben Taylor's Thinking Basketball, Dr. Lorena Martin's Sports Performance Measurement and Analytics, and Wayne Winston's Analytics Stories. These literatures can support this research. The book Analytic Method In Sports using mathematics and statistics to understand data from baseball, football, basketball and other sports. The main content around mathematics and statistic and probability methods and calculation in sports field. And all of data relative different sports athletics data. The book main idea about how to use statistics and probability to analysis sports data. By reading this book, I found the big value about application of mathematics in sports data analysis. Ben Taylor's Thinking Basketball provide some thinking logical about no-number sports analysis methods. Standing on basketball views, the book gives us a lot of insights on basketball analysis from where the data can't be seen. We can get some stories from this book data analysis. This book also tell us many analysis and strategy trap in basketball data analysis. The data needs to be interpreted correctly, and the correct is made based on the interpretation result. The book Sports Performance Measure tell us many data analysis of athletes' performance measure. I will reference basketball dunk analysis to know how to measure a basketball players performance. We should know we should choose different measurement methods and

measurement focuses according to different sports. We need to develop different analysis strategies according to sports. These reading material have different views of sports data analysis. They provide different views let us understanding sports measure world. But the they also have some disadvantages in sports data analysis. These literatures did not make a comprehensive interpretation of sports data analysis world from the perspective of an amateur. For example, when we analysis a basketball game, we are not only analysis team data, but also including individual player game data, player body conditions measure and analysis, and team playing strategies. These all belong to sports data analysis world. If we put all of these thing in this research, the readers will more easily understanding sports data analysis world.

Comparing with other literature, these literature are easy to understand, authoritative, and reliable. The documents contain a large amount of theoretical analysis and data analysis, which are very suitable for research topic on sports data analysis. These literatures use qualitative analysis and quantitative analysis

6.c Limitations of the Theory or Method of Research

The limitations of research method is that rely too much on literature and lack practical application. All research conclusions are based on the summary of the literature. During the research process, these was no more research with relevant professionals such as coaches or players. The theoretical research of sports data may be different form the way in actual work.

One of the research goals is to make people better understand the process of sports data analysis, but these is no guarantee that all fans have a theoretical basis in mathematics and statistics, and the content may cause difficulties for some readers. If the explanations of mathematics statistics theory is too basic, the research purpose of the article will be inconsistent with the content. Some sports data analysis theories may change with further research in the future. Because this research based on other literature research and make related conclusion. If the literatures change, the this research theory would be change.

6.d Recommendations or Suggestions of Future Academic Study

Continuing to explore the methods of statistics in sports data analysis.

Interview more sports related people to get advice and discussion on real sports data analysis and application.

Collect more samples, or use existing samples to combine data analysis techniques such as data visualization, machine earning and practice sports data analysis process.

Chapter 7

7.a Complete List of all Sources Used Regardless of Citation or Inclusion

Thomas, A. SEVERINI (2014). *Analytic Methods In Sports Using Mathematics and Statistics to understand Data from Baseball, Football, and Other Sports, CRC PRESS*

Ben, T. (2016). *Thinking Basketball*

Martin, L. (2016). *Sports Performance Measurement and Analytics The Science of Assessing Performance, Predicting Future Outcomes, Interpreting Statistical Models, and Evaluating the Market Value of Athletes, Pearson Education*

Winston, W. (2021) *Analytics Stories Using Data to Make Good Things Happen, John Wiley & Sons, Inc.*

How to Use Statistical Analysis When Betting on Sports. (n.d.). Gambling Sites.com.

https://bdtechtalks.com/2020/07/29/self-driving-tesla-car-deep-learning/

Dickson,B. (2020, July 29). Why deep learning won't give us level 5 self-driving cars. TechTalks. https://bdtechtalks.com/2020/07/29/self-driving-tesla-car-deep-learning/

The Role of Data Science in Sports.(2020, September 23). Master's In Data Science. Retrieved April 25, 2021, from https://www.mastersindatascience.org/resources/big-data-in-sports/

Kenton, W. (2020, November 27). Quantitative Analysis (QA). Investopedia. https://www.investopedia.com/terms/q/quantitativeanalysis.asp

Smith, T. (2021, April 22). Qualitative Analysis. Investopedia. https://www.investopedia.com/terms/q/qualitativeanalysis.asp

Australian Bureau of Statistics. (n.d.). Statistical Language – What are Variables?. https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language+-+what+are+variables

UCI Department of Statistics. (n.d.). What is Statistics. https://www.stat.uci.edu/what-is-statistics/

BYJU'S. (n.d.). Probability. https://byjus.com/maths/probability/

Complete Dissertation By Statistics Solutions. (n.d.). What is Linear Regression? https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/

SAS. (n.d.). Machine Learning What it is and why it matters.

https://www.sas.com/en_us/insights/analytics/machine-learning.html