UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Matemáticas y Computacionales

Using big data techniques to measure the performance of professional basketball teams

Trabajo de integración curricular presentado como requisito para la obtención del título de Ingeniero en Tecnologías de la Información.

Autor:

Corella Parra Brian Andrew

Tutor:

PhD. Cuenca Pauta Erick Eduardo

Urcuquí, junio 2021

Urcuquí, 4 de junio de 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES
CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN
ACTA DE DEFENSA No. UITEY-ITE-2021-00013-AD

A los 4 días del mes de junio de 2021, a las 14:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. FERNANDES CAMPOS, HUGO MIGUEL , Ph.D.
Miembro No Tutor	Dr. ANTON CASTRO , FRANCESC , Ph.D.
Tutor	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.

El(la) señor(ita) estudiante **CORELLA PARRA, BRIAN ANDREW**, con cédula de identidad No. **1729397982**, de la **ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES**, de la Carrera de **TECNOLOGÍAS DE LA INFORMACIÓN**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-43-No.496-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **Using big data techniques to measure the performance of professional basketball teams** , previa a la obtención del título de **INGENIERO/A EN TECNOLOGÍAS DE LA INFORMACIÓN**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Miembro Tribunal De Defensa	Dr. ANTON CASTRO , FRANCESC , Ph.D.	7,0
Presidente Tribunal De Defensa	Dr. FERNANDES CAMPOS, HUGO MIGUEL , Ph.D.	10,0
Tutor	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.	10,0

Lo que da un promedio de: **9 (Nueve punto Cero)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

Certifico que *en cumplimiento del Decreto Ejecutivo 1017 de 16 de marzo de 2020, la defensa de trabajo de titulación (o examen de grado modalidad teórico práctica) se realizó vía virtual, por lo que las firmas de los miembros del Tribunal de Defensa de Grado, constan en forma digital.*

CORELLA PARRA, BRIAN ANDREW
Estudiante

Firmado electrónicamente por:
BRIAN ANDREW CORELLA PARRA

HUGO MIGUEL FERNANDES CAMPOS

Assinado de forma digital por HUGO MIGUEL FERNANDES CAMPOS
Dados: 2021.06.07 15:54:14 -05'00'

Dr. FERNANDES CAMPOS, HUGO MIGUEL , Ph.D.
Presidente Tribunal de Defensa

Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.
Tutor

ERICK EDUARDO CUENCA PAUTA
Digitally signed by ERICK EDUARDO CUENCA PAUTA
Date: 2021.06.07 18:44:20 -05'00'

Dr. ANTON CASTRO , FRANCESC , Ph.D.
Miembro No Tutor

FRANCESC ANTON CASTRO
Signé électroniquement par FRANCESC ANTON CASTRO
cn=FRANCESC ANTON CASTRO, ou= ENTIDAD DE CERTIFICACION DE INFORMACION, o= SECURITY DATA S.A. 1, c= EC
Date: 2021.06.08 13:26:26 ECT

TORRES MONTALVÁN, TATIANA BEATRIZ
Secretario Ad-hoc

TATIANA BEATRIZ TORRES MONTALVAN
Firmado digitalmente por TATIANA BEATRIZ TORRES MONTALVAN
Fecha: 2021.06.07 14:35:43 -05'00'

Autoría

Yo, **Brian Andrew Corella Parra**, con cédula de identidad **1729397982**, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el autor del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Marzo del 2020.

Brian Andrew Corella Parra
CI: 1729397982

Autorización de publicación

Yo, **Brian Andrew Corella Parra**, con cédula de identidad **1729397982**, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Urcuquí, Marzo del 2020.

Firmado electrónicamente por:
BRIAN ANDREW CORELLA PARRA

Brian Andrew Corella Parra
CI: 1729397982

Dedication

"To the woman of my life and to those I hold dear in my heart. "

Acknowledgments

A special thanks to those who have supported me through thick and thin, always letting me know I am not alone and that everything is going to be ok. Especially to my one true love, Mika. I'm gonna be a superstar, watch me.

Resumen

El siguiente estudio trata sobre la analítica de big data en el baloncesto con el objetivo de introducirlo al país de Ecuador. Ecuador es uno de los muchos países latinoamericanos que carecen de una cultura de análisis de datos. Las organizaciones de baloncesto como la NBA tienen resultados notables al utilizar esto, ya que son la liga de baloncesto más grande del mundo, desde el punto de vista competitivo y del mercado. Además, los datos del baloncesto convencional todavía tienen un largo camino para cuantificar el juego en números. Relativamente, algunos eventos simples de baloncesto que no se cuantifican se cuantificarán para incluirlos dentro de la data. Los datos de baloncesto se obtienen de fuentes ecuatorianas para utilizarlos en el análisis de datos. Se utilizan varios métodos estadísticos de última generación para analizar los datos de baloncesto para demostrar los diversos resultados, conocimientos e interpretaciones que se pueden hacer con ello. Los resultados muestran muchas interpretaciones que los entrenadores pueden hacer cuando se presenta este trabajo, y se realizó con la ayuda de 2 expertos en baloncesto del país. La analítica de big data puede introducirse en el mundo del deporte en Ecuador para crear una liga más competitiva y objetiva donde los jugadores y entrenadores puedan aprender y sacar mucho provecho de mirar los datos y lo que tienen para ofrecer.

Palabras Clave: Analítica de Big Data, Baloncesto, Ecuador, Perspectiva

Abstract

The following study is about big data analytics in basketball, intending to introduce the concept to Ecuador. Ecuador is one of many Latin American countries to lack a data analysis culture. Basketball organizations like the NBA have remarkable results utilizing this concept as they are the most prominent basketball league globally, competitively, and market-wise. Also, conventional basketball data still has a long way to put the game into numbers completely. Relatively, simple basketball events will be quantified to be included as part of the data. Basketball data is obtained from Ecuadorian sources to use for data analysis. Various state-of-the-art statistical methods are used to analyze basketball data to demonstrate the different results and insights. Results show many interpretations that coaches can make when presented with the help of 2 basketball experts in the country. Big Data Analytics can be introduced to the sports world in Ecuador to create a more competitive and objective league where players and coaches can learn and take away from the data and what it offers.

Keywords: Big Data Analytics, Basketball, Ecuador, Insights

Contents

Glossary

box scores The box score in Basketball is the official record of the stats that occurred in a basketball game. It is printed on a sheet and contains information about both teams after the game.. 1

centers The center is usually the biggest or tallest member of the basketball team. In the NBA, many centers are 7 feet tall or taller. The center can be a big scorer, but also needs to be a strong rebounder and shot blocker. 8

contested jump shots Any jump shot where the closest defender is within 3.5 feet. 1

defensive efficacy Defensive efficacy relates to the percentage of points a team allows their opponents to score per 100 possessions . 9

defensive help Defensive help is an event occurring when a player leaves their position to help their teammate. 9

defensive switches A defensive switch is a strategy where players exchange their positions to prevent the opponent from gaining an advantage.. 9

degree of shot contest Degree of shot contest refers to the amount of defensive pressure a player puts on another player who is shooting. 9

field goal A field goal is any shot attempt from the offensive team. 7

forced and unforced turnovers A forced turnover occurs when the player or team on offense loses the ball due to the result of an event provoked by a defensive player. An unforced turnover occurs when the player or team on offense loses the ball due to their own mistake. 1

game clock Game clock refers to the official game time. Normally a FIBA basketball game lasts 40 minutes, with four quarters of 10 minutes, but the game clock is stopped when there are interruptions. 13

lineup Lineup is an expression to state the five players who are playing. 10

momentum In sports, psychological momentum has been defined as a bi-directional concept, affecting either the probability of winning or the probability of losing as a function of the outcome of the preceding event. 1

NBA The National Basketball Association is a professional basketball league located in North America. The league consists of 30 teams and features many of the best basketball players on the planet.. 1

offensive efficiency Offensive efficiency relates to the percentage of points a team can score per 100 possessions. 8

pace The number of possessions per 48(NBA) or 40 (FIBA) minutes for a team or player. 1

passes in the paint Any pass where the player receives the ball inside the 3-second lane. 1

pick and roll The pick and roll is a common and effective two-person offensive action involving an offensive player setting a screen for the player in possession of the Basketball. The screener will then roll towards the basket looking to receive a pass from the ball-handler. 11

play by play data This is a descriptive replay of a basketball game. This details all the events of both teams in real-time. . 16

playoff stage The playoff stage is taken after the regular season where teams are eliminated in a best of 3,5 or 7 game format. 7

plays A play is every event occurring in a basketball game.. 12

plus minus The point differential when a player or team is on the floor. 10

point guards The point guard is the team leader and play-caller on the basketball court. A point guard needs good ball-handling skills, passing skills, as well as strong leadership and decision-making skills. Traditionally basketball point guards were small, fast players, and this is still often the case. 8

points allowed The number of points a team allows their opponent to score. 9

power forwards The power forward on a basketball team is usually responsible for rebounding and scoring in the paint. A power forward should be big and strong and able to clear out some space under the basket. 8

pressure An organized basketball defense in which the team on defense pressures the opponent in an attempt to force a turnover. 9

regular season The regular season is the initial phase of a tournament typically consisting of various games to establish positioning for the next elimination phase. 7

shooting guards The shooting guard in Basketball has the primary responsibility of making long outside shots, including the three-point shot. The shooting guard also should be a good passer and able to help the point guard with the ball handling. 8

shot charts A shot chart is a chart of the layout of the basketball halfcourt, indicating the positions of the shots made and missed from the teams and players.. 16

shot selection Shot selection refers to the different shots taken from the free-throw line, 2 point shots from six different zones on the court, and 3 point shots from five positions on the court.. 11

small forwards Along with the shooting guard, the small forward is often the most versatile player on the basketball team. They should be able to help with ball handling, make an outside shot, and get rebounds. The small forward is often a great defensive player as well. 8

spacing The space between offensive players. Great spacing is when all offensive players are 15 - 18 feet from each other. 1, 13

team chemistry The best way to describe team chemistry is that it is a cumulative team attitude. In other words, it is the way the team, as a whole, feels about itself and its chances to succeed.. 1

transition Transition is used to describe the movement from offense to defense or defense to offense after a change of possession. 9

trap The trap is a defensive scheme where two defenders attempt to pressure the ball handler into a turnover. 11

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem statement

Basketball has become one of many sports that contain high amounts of data available for many different analyses and modeling [1, 2, 3, 4, 5, 6, 7]. These analyses and models typically revolve around understanding and improving the performance of teams and players. They utilize game data involving the team and individual variables that are quantified. As time passed, video recording systems were built to capture all players on the court. The video recollection allowed teams to extract high amounts of data. Detailed and unique statistics can measure individual and overall team performance. There are basic stats normally organized in box scores in every basketball game for both teams. This data can be beneficial for giving coaches and players an objective view of their performance [8]. As much as box scores can be useful, there are aspects of the game that are not part of the box score, such as forced and unforced turnovers, passes in the paint, number of contested jump shots, pace and many more [9]. There are also other not-so-easily quantifiable aspects, such as team chemistry, player injury impact, momentum, spacing, among others [9, 10, 4, 11]. Current literature is attempting to study and model these variables. Nonetheless, these aspects are part of the complete picture regarding a basketball game and a team's performance in a game or season.
Thus, the problem arises: should every basketball team utilize data and data analysis, even in its simplest form? Can they be used to help coaches and team staff with their decision-making of player transactions or coaching philosophies? Can they help in these aspects, or are the currently existing models enough to help a team find success in their respective leagues? Can advanced statistics provide more insight for basketball coaches and players? Will introducing these easily observable basketball events not yet quantified provide a clearer picture? For this, the most competitive league in Basketball, NBA, is used for exploring the importance of data and statistical analysis. Furthermore, the NBA is miles ahead of their data statistic usage and application regarding the basketball world. They will be the model to follow to apply these methods to the Ecuadorian basketball scene.

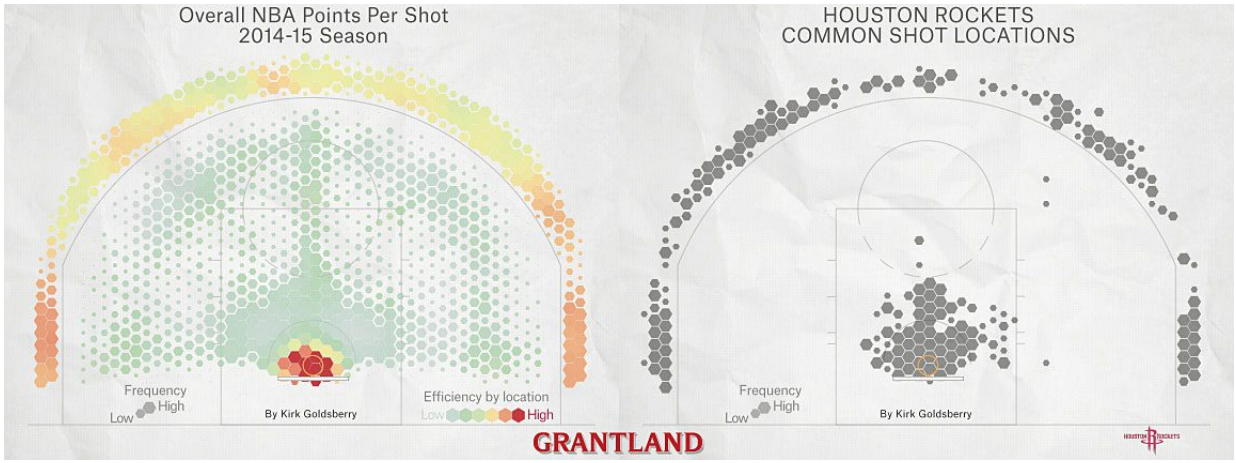

Figure 1.1: Shot Charts of the 2014-15 NBA season Overall vs Houston Rockets

The NBA commissioner Adam Silver stated that "Analytics are part and parcel of virtually everything we do now" at a conference in 2017[1]. Every NBA team has its team of data analysts working the data to find ways to get an edge over other teams. It all began in 2009 when the NBA implemented a state-of-the-art video system to track player and ball movements on the court. This video system allowed teams to collect enormous amounts of data to assess their performance in new ways[2]. Because of this, various aspects of the sport: from scouting rookie players, finding undervalued players, calculating efficient shots, the rise of the three-pointer, assessing player motor patterns for physical therapy, and team performance in specific events, can be performed with the aid of data analytics[3]. The general manager of the Houston Rockets at the time, Daryl Morey, a fellow computer scientist, was the first to push big data analytics to the game.

The play style of the Houston Rockets at the time of his tenure was mainly data-driven. His statistics team concluded that the best way to win is to maximize points per possession and thus find more ways to get more possessions, as he stated in a video by The Economist found on YouTube[4]. This team decided to maximize points per possession by shooting more three-point shots and only shooting inside the paint where there is a very high probability of scoring. This analysis would also attempt to bring players who best fit this philosophy and built a team able to uphold the data.

Figure 1.1[5] compares the Houston Rockets shots chart to the overall shot chart of the NBA in that season. The Rockets apply their interpretations of the big data analytics, shooting more threes and inside the paint, barely taking any mid-range shots. However, this trend of taking more three-point shots would spread across the NBA as more teams

[1] `https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116`[last access: February 4, 2021]

[2] `https://randerson112358.medium.com/how-data-transformed-the-nba-1cbc8b24e130`[last access: March 10, 2021]

[3] `https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116`[last access: February 4, 2021]

[4] `https://youtu.be/oUvvfHkXyOA`[last access: February 1, 2021]

[5] `https://digital.hbs.edu/platform-digit/submission/moreyball-the-houston-rockets-and-analytics/`[last access: January 29, 2021]

Figure 1.2: Bar graph displaying the increment of 3 point attempts from 1998 to 2017 in the NBA

began to take more three-point shots and increase their efficiency, as shown in the left part of the Figure. One of the biggest ways Big Data has already changed the NBA is in the importance of the three-point shot.

Figure 1.2[6] shows the increase of three-point shots taken per game for the last 20 years. This very simple, yet impactful interpretation has changed the NBA, as players efficient with shooting the three pointer became valuable. Data analytics would also impact the other facets of the game. For example, on defense, the team would encourage inefficient shots: a shot inside the point line and outside the painted area. Finding players who are skilled in shooting beyond the three-point line alongside others skills such as rebounding, steals, and assists became the norm for scouts. Players who are big and can shoot are now more sought after, as well as players who are skilled in defense and can shoot from long distance[7].

As big data analytics has played a role in changing the NBA, it will continue to change as the data evolves. NBA teams will adapt, and the data might lead to new interpretations and new changes[8]. Continuing the Houston Rockets story, the 2016-2017 season was their best, having the best record in the NBA and reaching the conference finals. However, heartbreakingly due to injury and poor decision making, the team lost the final game of the conference finals series to the eventual champions of that season. That game marked a new record, with the Rockets having the most consecutive missed three-pointers in a playoff game with 27. This event goes to show that the data cannot tell us everything. The Rockets shot poorly that game due to bad decision-making and bad shot attempts with no rhythm and movement. The data has shown to take more threes. However, the data cannot yet tell us how to create the best opportunities for the players to have a high chance of making the three-point shot against specific opponents, as that is up to the

[6] `https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116`[last access: February 4, 2021]

[7] `https://youtu.be/oUvvfHkXyOA`[last access: February 1, 2021]

[8] `https://www.theatlantic.com/entertainment/archive/2015/06/nba-data-analytics/396776/`[last access: January 28, 2021]

strategies of the coaching staff and the players' abilities. As such, the data comes in to support and help the staff in their strategic decisions. The more data available can lead to new interpretations, strategies, and solutions by the coaches to further improve their player development and team efficiency. In the NBA, the margin between winning and losing is often between the most minute details, which is why each basketball organization has its data analytic teams to process the data and find new ways to get an advantage over others[9].

Stepping away from the most prominent basketball league globally, basketball leagues in Europe are also finding their way inside the world of Basketball analytics. However, there are very few or almost no instances of this in Latin America. Two basketball experts, Prof. Manolo Albán and Paola Herrera, in Ecuador are consulted to provide further insight into the current situation of Basketball. In Ecuador, the Ecuadorian Basketball Federation (FEB) requires basic box score statistics for every official game. The coaching team usually tracked these standard box score stats manually to have the data. However, this changed in 2017, when the FEB implemented a statistics page on their website that would save this data with the help of a sports technology company: Genius Sports. Even so, most coaches would completely disregard this information or look at the numbers for their interpretations. There is no data culture in the country, as only one data analyst specialist works with Basketball currently in Ecuador. A news article by El Universo[10] mentions the 2021 Female Ecuadorian basketball champions having financial issues to compete in the South American Basketball tournament. Teams do not have the money to invest in people who can provide them with this service.

Many coaches are unaware of the benefits that data analysis can bring. The experts also claim that there is much talent to be exploited in Basketball. However, old-school coaches and players who are unwilling to evolve their philosophies to match successful basketball corporations continue to rule the basketball world in Ecuador. This concept can extend to Latin America. However, even if the basketball level in Ecuador and Latin America is not enough to compete with the more competitive leagues, introducing big data and advanced analytics to these countries could become a good start for the basketball world to become more competitive.

1.2 Objectives

1.2.1 General Objective

Measure the performance of professional basketball teams using big data techniques, including specific basketball events not usually included in the big data mining.

[9] `https://youtu.be/oUvvfHkXyOA`[last access: February 1, 2021]

[10] `https://www.eluniverso.com/deportes/otros-deportes/audaz-octubrino-no-solo-piensa-en-cuadrangul`
access: March 22, 2021]

1.2.2 Specific Objectives

- To introduce big data analytics using the Ecuadorian Basketball data and different statistical methods.
- To demonstrate the various types of data analysis and possible interpretations in the Basketball domain.
- To use big data analytics for determining the importance of the new Basketball variables.

1.3 Contributions

The contribution of this work is to perform data analytics with data from Ecuadorian Basketball. The results can provide objective evaluations for coaches and players alike that can improve and increase competition.

1.4 Document Organization

The following document is organized as such: Section 2 discusses the state-of-the-art stating the various scientific publications that have contributed to this topic and will be of utmost inspiration and motivation. Section 3 states the methodology of the thesis, exploring the data extraction and analysis phases. This section also provides insights into new basketball variables. Section 4 is about the results and discussion from the methods used in the previous section. Discussion relating to the ideas basketball coaches and players can have in this section. Finally, Section 5 ends the thesis with conclusions and future works to extend this thesis idea into even more relevance. The glossary explains basketball terms. The following links are used for the creation of the glossary[11] and also the following articles [5, 12]

[11] https://basketball.epicsports.com/basketball-glossary.html;//www.basketballforcoaches.com/basketball-terms/;https://www.ducksters.com/sports/basketballglossary.php;https://www.nba.com/stats/help/glossary/;https://www.rookieroad.com/basketball/stats/box-score/;https://www.sportsperformancebulletin.com/endurance-psychology/psychological-aides/role-momentum-sports-performance/

Chapter 2

State of the Art

This section presents the state of the art where statistical methods are used on different types and amounts of data to quantify basketball performance, provide new details for basketball events or variables, or predict future team performance. These methods can lead to different interpretations and new outputs such as new insightful metrics for player performance, models for future game predictions, finding the variables that differentiate winning teams from losing teams.

2.1 Performance in Basketball

Before big data analytics became significant, the coach would make decisions based on their knowledge, game philosophy, and experience. They would observe their team performance with the basic box score stats and make adjustments. Most of these coaches would evaluate their players individually, ranking their players from best to worst[1]. Everything was established based on the freedom of the coaches [8]. There was no metric to objectively evaluate players or teams besides the subjective observations and opinions of the team's coaching staff. In the big leagues, where Basketball is a business, and every team is financially responsible for their players, coaches, and other staff, these decisions could affect the team's financial resources [13]. As such, the team's leaders had to make financial decisions based on the knowledge and opinion of others [8].

Nowadays, as more data is available, big data analytics have allowed the creation of advanced basketball game statistic analysis that attempts to evaluate all aspects of the sport [7]. These advanced basketball statistics permit teams to objectively value players and helps them to consider their financial resources when making a decision.

According to Alamar [1], there are two data statistic goals. The first one is to save time regarding making the most efficient decisions by evaluating teams and players. The second one is to receive comprehensive data about players and teams that lead to informative insights that would not be possible without the data. Thus, there usually are two big reasons that justify big data analytics in Basketball and any sport. One is to thoroughly evaluate teams and players to improve present and future performance and aid financial

[1] `https://www.espn.com/nba/story/_/id/9980160/nba-how-analytics-movement-evolved-nba`[last access: February 11, 2021]

decisions of player transactions. As the data has become more available and scalable worldwide, there is a vast amount of scientific literature that tackles this subject of data analytics in Basketball. It offers new interpretations and results to broaden the scope of what big data analytics can do in this sport.

2.2 Using Statistical Analysis in Basketball

In 2003, Dean Oliver published a book entitled "Basketball on Paper: Rules and Tools for Performance Analysis" [9]. This book became the pioneer for introducing a more statistical perspective of the data to evaluate performance in Basketball, taking data from the '90s and early 2000s. Oliver also stated new interpretations about efficiency and winning. Oliver talks about advanced analytics is necessary to measure the genuine value of a player in the market. This book focused more on teams first and players second, as it is a team-oriented sport. It introduced the infamous four factors that break down winning efficiency. These four factors include shooting efficiency, rebounding percentage, turnovers per possession, and free throw attempts. They highlight Oliver's approach of using available data to suggest new ways to influence the game [9]. This book would introduce the American public to new strategies and ideas to improve their basketball teams.

2.2.1 Finding Game Performance Indicators using Box-Score data

Also, besides the NBA, studies were being performed in Europe, as Sampaio et al. [14] performed a statistical analysis on data from the Portuguese Professional Basketball league for the years 1997-1998 & 1998-1999 to attempt to find game statistics that differentiated winning from losing teams. In total, they used 353 regular season games and 56 playoff games. The authors divided the games by game location (home or away) and game category for each game type. The authors utilized a k-means clustering algorithm to group up the games according to various final score differences: close games are games that ended in a score difference of 1 to 8 points, balanced games from 9 to 17 points, and unbalanced games of above 18 points. The game statistics that were retrieved were 2 - 3 point field goals made and attempted, free throws made and attempted, defensive and offensive rebounds, blocks, assists, fouls, steals, and turnovers. The data was normalized by game rhythm to access team performance across the whole season. A discriminant analysis was applied to determine the game statistics to predict game outcomes after categorizing the data accordingly. After performing these analyses, for unbalanced games, losing teams performed poorly in all game statistics. However, in regular-season close games, home teams were more efficient from the 2 point line, made more free throws, and committed fewer fouls. Away teams shot fewer three-pointers, made more free throws, and caught more defensive rebounds. For the playoff stage, home winning teams committed fewer fouls and obtained fewer offensive rebounds, while away teams made more free throws and got more offensive rebounds. Playoff games also displayed fewer points, and players committed more fouls, which led to an increase in free-throws attempted.

A similar study was conducted by Csataljay et al. [15] which tries to distinguish game performance indicators from winners and losers in close games, as they most often exhibit the smallest of differences in performance. Data from 54 matches were collected

from the official score sheets of the 2007 European Basketball Championship. The game statistics used were 2 and 3 point attempts, makes and percentage, free throw attempts, makes and percentage, offensive, and defensive rebounds, total rebounds, assists, fouls, steals, turnovers, blocks, and total points. Cluster analysis was performed to group the games based on their point differentials into close (from 1-9 points), balanced (10 to 22), and unbalanced (above 22). They used Wilcoxon signed-ranks tests to compare the 18 game statistics between winning and losing teams. The winning teams had fewer 3 point attempts, higher shooting percentage, more defensive rebounds, and more free throws made, which also means increased free throw percentage. For balanced games, better shooting percentage, defensive rebounding, and assists were the difference makers.

This study focuses on finding game statistics that are performance indicators of winning teams playing on their home court vs. winning teams playing in foreign territories. De Rose[6] utilizes official statistics data from the Division I Men championship of Sao Paulo, Brazil, with 606 matches. The following game statistics were used: 2 and 3 point field goal made, attempted, and efficiency, free throws made, attempted, and efficiency, offensive, and defensive rebounds, total rebounds, assists, steals, blocks, turnovers, fouls, total points made, total of possible points, and total points efficiency. They used an ANOVA to find statistical differences in any of the variables. The results yielded favorably for the home teams, being statistically superior to the away teams in 3 pointers, defensive rebounds, assists, blocks, steals, turnovers, and total points efficiency. An essential remark from the article's discussion states, "It (statistical analysis) must be associated to another kind of qualitative observations and physical assessment. Combining all these observations will allow coaches and athletes to do a precise and consistent analysis of the situations to provide a better individual and team's performance" [6].

Similar to the previous study, there is an article by Sampaio et al. aimed at identifying the performance of the various playing positions with regards to game location being at home or away [16]. This study identified the game statistics that differentiate different player position performances of home and away teams. They utilized data from the 2004-2005 Euroleague with 225 games. The players were grouped up according to their positions: guards being the point guards and shooting guards, forwards being small forwards and power forwards, and centers as their group. The game statistics were 2 and 3 point field goals made and attempted, free throws made and attempted, defensive and offensive rebounds, blocks, assists, fouls, steals, turnovers, and minutes played. They performed a MANOVA with a discriminant analysis performed after. The success of the guards depended on 2 point field goals made, defensive rebounds, assists, steals, blocks, and committed fouls. For forwards, made free throws, assists, steals, blocks, and fouls were the essential stats that defined their performance. The analysis showed no significant differences in the center's performances for home and away games.

All these studies utilize data from one season or tournament year. However, this study by Ibanez et al. [17] focused on obtaining game statistics for indicating season-long performance success. They used data from the 2000-2001 and 2005-2006 regular Spanish season consisting of 870 games. The following game statistics were used: 2 and 3 point field goal made and attempted, free throws made and attempted, defensive and offensive rebounds, assists, steals, turnovers, blocks, and fouls. They controlled game rhythm and calculated possessions for getting the offensive efficiency of each team. Teams were classified as best teams if they made it to the playoff series, and those who did not make it were the worst

teams. They performed A one-way variance analysis to find the game statistics of best and worst team performances. Then a discriminant analysis was done to locate the game-changing variables among them. The best teams made more free throws, caught more defensive rebounds, had more assists, steals, blocks, offensive efficiency, and fewer fouls over different seasons. Thus, for a season-long success, assists, blocks, and steals were the most important statistics that teams performed better than their opponents.

Most of the above studies talk about general or offensive performance indicators. A more difficult approach is to attempt to find defensive indicators. However, Álvarez et al. [12] conducted a study of this topic in peak basketball using data from the 2008 Olympic basketball games. As their primary focus was on defense, they defined variables regarding the teams' defensive schemes and their success with them regarding wins and losses. The variables defined were type of defense, pressure in offensive transition, defensive switches, defensive help, passes in the paint, degree of shot contest, points allowed, defensive efficacy, and final result of the game. The authors conducted observational methods of all half-court game phases. In this highly competitive tournament, a quarter man to man defense was highly used and worked more than 53% of the time. Half-court zone, pressure, defensive switch, help, medium opposition provided the highest efficacy, meaning these strategies had the highest percentage of not allowing their opponent to score in that possession. However, usage rate should also be considered. Only 3% of the time, a team used a half-court zone. These usage rates could also state the factor into the efficacy results. Meaning teams only used this strategy in specific situations that resulted in positive outcomes.

2.2.2 Building models for predicting team performance

Finding and interpreting game statistic performance indicators can then attempt to build models that predict teams' performance in the future. Li et al. [13] proposed a data-driven performance prediction model from NBA data, similar to ESPN. They address optimizing a specific team's future performance in the competition with other teams by choosing players, determining their playing time in the court, and eventually maximizing possible winning probability for the next regular season. They use the non-parametric Data Envelopment Analysis (DEA) method as an ideal production frontier approach. The prediction process can be divided into two steps: the first step is to conduct a multivariate statistics regression analysis to estimate the quantity relationship between the winning probability and the team's various game outputs obtained in games. By taking the regression equation obtained in the first step as the objective function, the other step conducts a DEA-based player portfolio efficiency analysis to optimally choose players and allocate the playing time among players in the court. Afterward, the proposed DEA-based data-driven prediction model is applied to the data of an NBA team, the Golden State Warriors, from the four seasons of 2011-2015, attempting to predict their performance for the 2015-2016 season. This approach has various implications: no player injuries so that coaches can allocate all players with playing time and play in the next season. In the end, their prediction was extremely close to the prediction results provided by ESPN and FiveThirtyEight, with the model only having almost one more win for the respective season. They predicted the GSW to win 61 games in the 2014-2015 season.

More studies are being done on finding more accurate ways to predict team performances. A regression tree model by Huang et al. [18] was used for game prediction. They

used the Golden State Warriors for their analysis. The authors extracted data from the basketball-reference website that contains all statistical data regarding the NBA. The following variables were chosen for their predictive model: Games started, minutes played, 2 and 3 field goal attempts, makes and percentages, free throw attempts, and percentage, offensive and defensive rebounds, total rebounds, assists, steals, blocks, turnovers, personal fouls. The data was divided based on the season game. The first 50 games become the training set, the 51-66th games become the validation, and the 67-82nd games are the test sets. Regression tree, linear regression, and support vector regression models are trained and used to predict the validation dataset having the root mean square error to determine the better model, predicting player scores using the test set. The predicted team score is computed by summing the predicted individual player scores. By comparing both team's predicted scores, the model can provide a predicted outcome of the game. The prediction accuracy was 87.5%, correctly predicting the number of points for the Warriors and their opponents in certain games. Thus, the authors correctly state that the regression tree model can indeed effectively predict team scores. The limitations of this study include avoiding factors such as player injuries or load management. More teams should be included to verify the regression tree model's ability to predict game scores.

More studies try to remodel or adjust already existing statistics to consider more actions in Basketball. This article by Grassetti et al. [19] creates an adjusted plus minus framework. This study replaces the classical response variable (scored points) with a more comprehensive score that combines a set of box score statistics and focuses on team lineups instead of individual players. 240 matches from the 2018/2019 regular season of the Turkish Euroleague Championship, regarding play-by-play and box scores, were used. The model used to evaluate performance efficiency is a Bayesian model. This model considers entire five-player lineups, and the interaction of the players is measured considering the entire lineup. Measuring lineups based on their effect on the court or their team's overall rank is done in this study and can lead to different interpretations. This model gives new insight into how data analytics should be performed in the future, focusing on the team rather than individual data.

2.3 Using Machine Learning Algorithms for various aspects of Basketball using Big Data

Machine learning algorithms are also used for traditional studies like those above regarding finding the game statistic performance indicators in a more detailed manner. Migliorati [20] uses box scores from 2004-2005 to 2017-2018 NBA seasons of the Golden State Warriors. Dean Oliver's four factors of success [9] are used as classification independent variables. The authors perform Classification and Regression Tree (CART) and Random Forests in this article. They obtain a CART box score model and eliminate the points and assists stats because they do not provide further insight. For the GSW, the tree shows that they win when the team makes at least 10 3 point shots and catches more than 30 defensive rebounds. The tree also shows the team loses if they fail to make 10 three-pointers and allow the opponents to catch more than 35 defensive rebounds. The variable importance of the CART box score reveals that defense is the critical factor for this team's success.

The authors then perform this same type of analysis with the four factors of basketball performance. The random forest also complements the CART models, stating how the defense is vital for winning basketball games. A 71.68% accuracy is obtained for the CART box score model, and a 67.26% accuracy for the CART four factors model with the absence of shooting statistics. For the random forest, a 90% accuracy is obtained and only a 94% accuracy with the 4 factors, including shooting. This analysis can be done in other basketball leagues.

Machine learning is not only limited to the former types of studies; there is a study by Ivanković et al. [2] that uses data mining by neural networks to determine how shot selection influences game outcomes. This study touches a more specific area of the data. The authors include more variables that also influence the outcome of a match, and they include offensive and defensive rebounds, assists, steals, turnovers, and blocks. This paper uses the Serbian men's First B Basketball league for data mining, having all games from five seasons: 2005-06 to 2009-2010. Neural networks perform data mining analysis. They computed the shot percentages of each position first and used them as input parameters for the network. The neural network outputs one parameter, which is the result of the match: win or loss. Once the network has been trained, separating the data into training and test sets, a relation is obtained determining how the different inputs influence the game. This network yielded only a 66.4% accuracy, making the authors include new data into the model. So they took data regarding offensive and defensive rebounds, assists, steals, turnovers, and blocks and performed the same procedure to retrain the network and obtain the influences of these parameters on the game results. This new model accurately predicts game results 80.96%.

Another use of machine learning is with the study by Tian et al. [4]. Most studies and sports analytic systems focus on the offensive side of the game. The defensive side is also 50% of the game, so this study uses machine learning to classify different defensive strategies using player and ball tracking datasets from the NBA using SportVU tracking data. The data obtained was from the 2012-2013 NBA season, from about 630 games using 32,377 possessions. Three machine learning models, K-Nearest Neighbor, Decision Trees, and Support Vector Machines were used to identify a defensive strategy commonly used against a pick and roll, which is the switch and trap. The model uses three different attributes to classify this defensive strategy: location data at each time step, the distance between players, and defensive zone. With the model's specifications, the authors then used 10-fold cross-validation to train and validate the classification model, having a manually labeled dataset. Each machine learning algorithm took two approaches: location data and the other using movement vectors and distance. A higher accuracy percentage was obtained by the algorithms using the second approach. Among the three machine learning algorithms, KNN yielded the best accuracy. However, the algorithm can be improved by using a player-specific dataset to strengthen the system's ability to define each defensive player and determine the defensive relationship between players. This study can extend to various offense schemes utilized by teams other than the pick and roll and extend to different defensive schemes.

Basketball data analytics can be particular as this study uses recurrent neural networks (RNN) to predict the success of a three-point shot. Shah et al. [10] use NBA SportVu dataset of over 20,000 three-point shots and compares the RNN with a static feature-rich machine learning model using angle and velocity. The RNN architecture contains two-

layered long-short-term memory (LSTM) units using peephole connections. The input to the LSTM is data of the ball location in the three dimensions over time and the game clock. The RNN predicts the possibility of a made shot and the parameters for the mixture density network (MDN), consisting of three mixtures of tri-variate Gaussians. This RNN model produced the highest classification scores. By only using positional data, it easily outperformed the static machine learning models. The study suggests that RNNs can learn sequential behavior. Since RNN can be utilized, improving hyper-parameters is the first step to optimize this model thoroughly.

2.4 Performing Advanced Data Analytics on qualitative aspects of Basketball using Big Data

As mentioned before, the introduction to data analytics in Basketball has become huge in decision making and improving performance in the short or long term [9]. However, from a practical standpoint, this data is far too superficial or straightforward to give more complete conclusions about the game of basketball [8]. Specific data is required for these types of analyses, such as play-by-play and video tracking data.

One of the many things that are not quantified publicly is team chemistry on the court. Maymin et al. [11] introduced a Skills Plus Minus (SPM) framework to measure on-court chemistry regarding basketball skills. They chose three basketball skills: scoring, rebounding, and ball-handling to measure on-court chemistry. The theoretical framework attempts to find synergies between skills, decompose players' contributions according to these skills, find synergies of teams regarding these skills, provide context-dependent player ratings, and find mutually beneficial trades between all teams. Play-by-play data from 2006-2007 to 2009-2010 NBA seasons includes 4718 games and 987,343 plays. Their model is a series of nested probit regressions that predicts the likelihood of various events for a given play. This framework stimulates an entire basketball game given the estimated coefficients to predict an event and the resultant end of play variables taking all ten players into account, introducing a probability of events regarding lineups and specific players. This framework gives players like LeBron James a high-scoring skill, meaning a team with LeBron included in the lineup outscores their opponents by 15.2 points. It also results in another player, Chris Paul, being the best ball-handling player by outscoring their opponents by 4.8 points regarding this skill. The framework states a positive synergy of offensive ball handling with offensive rebounding and scoring, meaning players who handle the ball well will often play better with teammates who give them more chances at scoring and with teammates who are good at scoring. Also, players with the same skillset can result in negative synergy, as offensive scoring can hinder. This framework is unique and can be utilized for future research, such as finding the optimal distribution of minutes for each team player to maximize their performance and synergistic lineups or create a synergy factor that attempts to improve' skills while playing with specific players.

Teammate interaction is also a fundamental component of team sports, as players work together with one main goal in mind [9]. Sandri et al. [21] investigate the relationships between each player's performance variance and the team lineup composition. It does this by assuming shot-varying transition probabilities between regimes, highlighting the positive

and negative interactions between teammates. The authors used play-by-play data from the Golden State Warriors for the 2017-2018 NBA regular season to develop an actual data case study for this dynamic model. Markov switching models with time-varying transition probabilities were used to model shooting performance, as it is subject to improvement or worsening over time. A network graph can be obtained with these parameters to visualize players' interactions and whether these interactions effectively translate into team performance. To do this, they defined a score for these interactions and also measured the points scored by these lineups. They found a direct relationship as the interaction score and points scored are mutual. The results can be of utmost help for coaches when they attempt to find the best lineups and pairs of players that work best together. A limitation of this model is that it cannot be used with little data.

Basketball is also meticulous for spacing concept, which is also often disregarded in the conventional box-score [9]. Metulini et al. [22] attacked this issue by using cluster analysis on Spatio-temporal trajectories extracted from a GPS tracking system to characterize players' positions on the court. A player's position on the court depends on the teammates' and opponents' position and is most likely due to some previous strategy. They had a friendly basketball game with six players who wore microchips on their clothing. The players' positions were recorded every 161 milliseconds, based on their x, y, and z coordinates. The system recorded a total of 133,662 space-time observations. The data was processed using a Kalman approach, filtered by dropping the data unrelated to time playing on the court, such as pre-match, half-time, and post-match. K-means clustering analysis was performed to group player positions regarding time instants. There were 8 clusters that each represented different spacing and movement patterns. 4 of them were found to be offensive possessions, while another was a defensive event. For future studies, this analysis can be done regarding multiple matches, matching this data with a play-by-play to obtain further insights into the team's performance regarding specific spatial patterns taken.

The previous study is extended by Metulini et al. [23], where they model the pattern of surface area in the basketball court and measure its effect on the team's performance. Player coordinates from three matches in 2017 taken from the Italian professional basketball league. The players wore microchips that collected their x,y, and z positions on the court. They filtered instances where they only had information when the game clock was running. A three-step procedure was proposed. A Markov Switching Model is used to detect structural changes in the surface area, followed by a descriptive analysis to extract associations between game variables. Then a vector auto-regressive model was used to assess the relationship between state probabilities and scored points. The surface area drastically changes when going from an offensive position to a defensive position. The surface area tends to be narrow or smaller in defense than the large surface areas found in the offense. Further studies include introducing ball trajectory as part of the data to visualize its impact on surface area, spacing, and performance.

Frequently, there are high-pressure situations where a specific moment of the game requires players to level up or show the best version of themselves to get ahead of their opponents or catch up. It is important to note that high pressure concerns game situations and events and is not regarded with psychology. There could be internal factors for players to express themselves as being stressed. However, an external factor such as they need to score 3 points is the high-pressure example for this section. There is a study by Zuccolotto et al. [5] that analyzes shooting performances in these pressure situations. They utilize

play-by-play data from the Italian Serie A2 championship of 2015-2016 to build and validate the model using data from the 2016 Olympic basketball tournament. The dataset yielded over 70,000 shots. They had to estimate how scoring probability changes due to different game situations. They define four events for a high-pressure situation with the help of basketball experts. These are the expiring time of the shot clock, point differential is minimal between the teams, poor performance by the team up to that moment, and when one player has missed a previous shot. They perform multivariate analysis for all these events for the different types of shots and then perform a CART to find relationships between these situations and the shots. They found the percentage of field goals decreased as the shot clock was closer to zero and also when a player has missed a previous shot. Similar results were validated with the Rio dataset, meaning that this also happens in higher competitive leagues. They also show a new metric that distinguishes players from their propensities to shoot in certain situations and their performance on shooting and overall team performances. These can allow coaches and players to identify personal reactions to these situations. This study can be furthered by having data regarding the distance of the shots and distance of defenders when attempting a shot.

2.5 Discussion

The basketball state-of-the-art has a lot to offer. Table 2.1 provides a summarized list of the applications with the different statistical methods. As more advanced data apart from the usual box score data is obtained through video and tracking systems, models for prediction and performance can become more complex and complete utilizing different machine learning algorithms. For example, cluster analysis is used in many studies for separating games according to point differentials. However, cluster analysis is used for different purposes when using live-data or play-by-play data, such as separating games according to the spatial relations of players on the court or classifying defensive events. Discriminant, regression, and ANOVA are used with box-score game data to find performance indicators that distinguish winning teams from losing teams.

This can extend to various ideas, such as finding game variables that differentiate winners in home games and away games using MANOVA. A Neural network was also used to find game performance indicators. However, when given live video data extractable from a video system software, it was also used for predicting the outcome of a 3 point shot. A DEA approach was used to model future prediction using player totals. Bayesian Analysis was compared with regression analysis for measuring player lineup efficiencies using play-by-play data.

Nested Probit Regression was performed with play-by-play data to model a basketball game to create a new skilled index for various applications. Markov switch models utilize live-video data to classify different surface areas that teams create while playing on the court and use play-by-play data to model shooting variability depending on the player lineups that provide an insight into teammate interactions. Machine Learning techniques are also used with live video data. SVM, Decision Trees, KNN are put to the best to identify the best classifier of defensive events. CART and Random Forest use complete box-score statistics to model and visualize team performance for future predictions. Regression trees are also used to compare against support vector regression and linear regression to de-

termine the better model. Many different applications can be performed by statistical methods that can also depend on the available type of data. Table 2.1 summarizes the applications and statistical methods found in the state-of-the-art.

Application	Statistical Method
Separate games according to point differential	K-means Clustering [14, 15]
Determining performance indicators of professional basketball teams	Discriminant Analysis [14, 16, 17], Factor Analysis, One-Way ANOVA [6, 17], Wilcoxon signed-ranks tests [15]
Estimating the relationship between game variables and winning probability	Multivariate Regression Analysis [13]
Determining the influence of shot selection on the outcome of a game	Feedforward Neural Network[2]
Classifying the correct defensive event using location data or movement vectors and distances	KNN [4], Decision Trees [4], Support Vector Machines [4]
To model future predictions	Bayesian Analysis [19], Data Envelopment Analysis [13]
Finding significant differences between game location performances and player roles	MANOVA [16], Discriminant Analysis [16]
Estimate the likelihood of events for a particular play in Basketball	Nested Probit Regression [11]
Creating a model for classifying a team's match performance	CART [20], Random Forests [20]
Determine structural changes in surface areas for team movements on the court	Markov Switching Models [23]
Model shooting performance	Markov switching Models [21]
Estimating the relationships between game variables to obtain a player score	Support Vector Regression [18], Regression Tree [18], linear regression [18]
Predicting the event of a 3 point shot	Recurrent Neural Network [10]
Characterize players' positions on the court	Clustering [22]
Estimate scoring probability changes from different high pressure game scenarios	Multivariate Analysis [5]
Find relationships between the different shots and high-pressure situations	CART [5]

Table 2.1: State-of-the-Art Statistical Methods used for various Basketball applications

Chapter 3

Methodology

This section presents information regarding the type of data extracted and where it is from alongside the methods for data analysis with the proper justification. Following the state-of-the-art, there are many paths to take, but it all depends on the type of data available. NBA statistics are the inspiration. The NBA is the best, and biggest Basketball league worldwide due to its market and competition level [13]. They are already hundreds of kilometers ahead regarding statistical analysis. It makes sense to follow or catch up to what they have been doing, as it works for them. It serves as a basis for others to use and start the journey.

3.1 Phases of the Methodology

Figure 3.1 provides an outline of the steps taken in this analysis. There are two sections, one regarding data and the other section about analysis of this data.

3.1.1 Data Phase

The only available basketball game data from Ecuador can be found in the Ecuadorian Federation of Basketball (FEB) website[1]. This website has a statistics page with actual tournament-related data such as box-score data, play by play data, and shot charts from individual games, from various tournaments from 2017 to 2020. This website is powered by Genius Sports Group, a company dedicated to providing technology all over the sports world for data collection and analysis. However, there has a significant downside. This information is not available for immediate download for usage as other websites such as basketball-reference, a website for all sports data of the US. Manual extraction had to be done. Also, because of this, only the box-score information can be easily extracted. As data from previous years is discontinuous, meaning different teams and players change over the years, this is another limitation as information per year would not be logical. There would be data from different teams and players, and a discontinuity will be formed.

[1] `http://ligas.feb.ec/competitions/?cu=FECUB/standings`[last access: April 6, 2021]

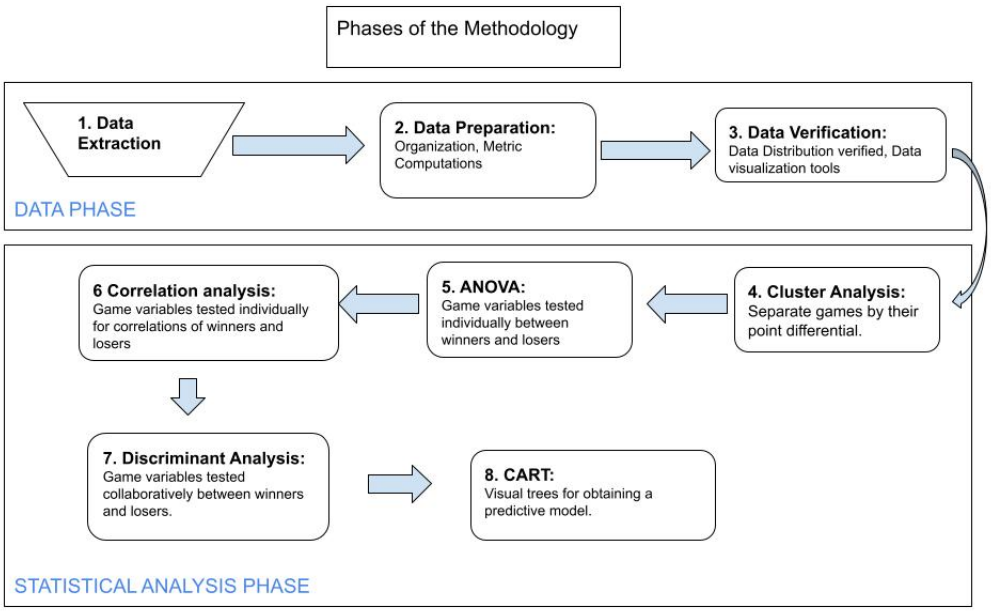

Figure 3.1: Phases of the Methodology

Season	Team	W	L	X2p	X2pa	X3p	X3pa	ft	fta	orb	drb	ast	stl	blk	tovs	pf	pts
FEM2020	Victoria Cogarol	7	0	128	328	50	203	94	157	123	206	79	108	7	141	128	500
FEM2020	Audaz Octubrino	6	1	127	301	23	95	75	127	57	192	83	98	5	146	93	398
FEM2020	La Perla S.C.	3	4	85	254	34	148	64	128	53	216	50	57	15	164	118	336
FEM2020	UDJ	2	5	85	252	43	144	97	166	50	231	75	64	5	168	145	396
FEM2020	Deportivo Cuenca	2	5	105	294	38	159	47	76	83	237	76	66	10	160	97	364
FEM2020	VO4	1	5	80	251	29	132	66	122	66	174	55	66	13	132	102	313

Figure 3.2: Standard Tournament totals for every team of the 2020 Female Basketball Tournament

For this reason, box-score totals were obtained from the Official Female Ecuadorian Basketball (FEB) tournament that was postponed in 2020 (due to the COVID-19 pandemic) and finished in January of 2021. This league consisted of six teams from all across the country. The format had a regular-season phase where teams played a total of seven games each, followed by an elimination stage where the four best teams were divided into semi-final stage consisting of a best of 3 series, having the victors of those series then face off in the finals similarly in a best of 3 series. The data consists of 28 total basketball games, 21 from the group stage and 7 from the elimination stage. Finally, as there is no state-of-the-art video system being implemented in the Ecuadorian Federation of Basketball as in the US and Europe, there is no video data extractable besides live recordings of some basketball games the FEB's Facebook page.

First, manual extraction of all box scores from all group stage games was performed. The box score is extracted according to a specific order: minutes played, total field-goals made, total field-goals attempted, field-goal percentage, 3 point shots made, 3 point shots attempted, 3 point percentage, free throws made, free throws attempted, free throw percentage, offensive rebounds, defensive rebounds, total rebounds, assists, steals, blocks, turnovers, personal fouls, points, and then an additional / metric. Table 3.1 explains these stats. This order is more organized as the shooting category is first, made and attempted, then percentages for visual purposes. Then non-shooting offensive categories are organized by rebounds, assists followed by defensive categories like the steals and blocks, then by variables that generally have a negative impact, such as turnovers and personal fouls, followed by points. This order serves as a better aesthetic to the box score. The box-score is found in Figure 3.2.

Box-Score Statistics	Description
Players	The name of every individual player part of the team, regardless of whether they participated in the current match.
MP (Minutes Played)	The amount of time in mm:ss that the player was on the court.
FG (Field Goals)	The number of made 2 point shots from anywhere inside the 3 point line.
FGA (Field Goal Attempt)	The number of attempted 2 point shots from anywhere inside the 3 point line.
FG%(Field Goal percentage)	The rate of field goals made per field attempted. Calculated as FG/FGA.
3P (3 point field goals)	The number of shots made behind the three-point line.
3PA (3 point field goal attempts)	The number of shots attempted behind the three-point line.
3P% (3 point field goal percentage)	The rate of 3 points field goals made per 3 point field goal attempt. Calculated as 3P/3PA.
FT (Free Throws)	The number of free-throws made from the free-throw line after suffering a shooting foul.

FTA (Free Throw Attempts)	The number of free throw attempts from the free-throw line after suffering a shooting foul.
FT%(Free Throw percentage)	The rate of free-throws made per free throw attempts. Calculated as FT/FTA.
ORB (Offensive Rebounds)	The number of rebounds when on an offensive possession.
DRB (Defensive Rebounds)	The number of rebounds when on a defensive possession.
TRB (Total Rebounds)	The number of total rebounds that includes defensive and offensive. Calculated as ORB + DRB.
AST (Assists)	The number of passes directly led to a made field goal from anywhere on the court by any teammate.
STL (Steals)	The number of times the player took the ball away from their opponent.
BLK (Blocks)	The number of times the player legally deflected an opponent's field goal attempt.
TOV (Turnovers)	The number of times the player lost the ball, whether forced or unforced. Such as making a bad pass out of bounds or getting the ball stolen from him.
PF (Personal Fouls)	The number of times the player commits an illegal action against his opponent. Such as pushing an opponent, hitting his arm, and many more events.
PTS (Points)	The number of points scored by the player. This is calculated by FT + 2*FG + 3*3P.

Table 3.1: Box-Score Statistics with their respective descriptions

After data extraction for each team in every individual game, tournament totals are extracted for the group stage for every team to obtain a tournament average. The final data included the variables Standings, team name, wins, losses, points, points allowed, assists, defensive rebounds, offensive rebounds, steals, turnovers, 2 point shots attempted, 2 point shots made, 3 point shots attempted, 3 point shots made, free throws attempted, free throws made, followed by bench points, fastbreak points, 2nd chance points, points in the paint, points off turnovers, blocks, minutes played, and personal fouls. Luckily, the FEB website provides bench points, fastbreak points, 2nd chance points, points in the paint, points off turnovers in their match summary. These are normally extracted manually by using play-by-play data, live analysis, or game recordings. Shooting efficiency was calculated per team, such as FG%, 3P%, and FT%. The percentage of points per shot was also calculated. These additional statistics are described in Table 3.2.

Extended Statistics	Description

_PG	Any statistic per game. For example PPG: Points per game, APG: Assists per game, etc.
BP (Bench Points)	The number of points scored by the players who were substituted at any point in the game.
FBP (Fastbreak Points)	The number of points scored when entering a fastbreak event. This event is the rapid action by one team to go to the other side of the court to gain an advantage in order to score a basket, generally within 6 seconds of possession.
2CP (2nd Chance Points)	The number of points scored after getting an offensive rebound, hence getting a 2nd chance to score on the same possession.
PPNT (Points in the Paint)	The number of points scored within the restricted area or paint.
PTOV (Points off Turnovers)	The number of points scored off an opponents turnover.
%ofFG (Field Goal contribution)	The percentage of total points being scored by field goals. %ofFG = (FG*2)/PTS.
%of3P (3 Point field goal contribution)	The percentage of total points scored by 3 point field goals. %of3P = (3P*3)/PTS.
%ofFT (Free Throw contribution)	The percentage of total points scored by Free throws. %ofFT = FT/PTS.

Table 3.2: Additional Stats found in FEB summary or easily computable.

After the information was extracted, totals were divided by the number of games to set an average value per game. This value is good for visualizing the average performance of a team per game in this specific tournament. An extract of this data is found in Figure 3.2. Later, some metrics can be extracted with this data that provides a better insight into performance. These metrics are used in the NBA by statisticians to present to the public and creates a means for analysts to discuss in their shows. These are Effective Field Goal Percentage (eFG), Turnover percentage, Offensive rebounding percentage, and Free throw rate. These come from the four factors of basketball success by Dean Oliver [9] used by NBA statisticians to this day. They are described in Table 3.3

Advanced Box-Score Statistics	Description
eFG% (Effective Field-Goal Percentage)	This metric adjusts the extra point obtained when making a 3 point shot. It is computed as: eFG% =(FG + 1.5*3P)/(FGA+3PA)
TOV% (Turnover Percentage)	An estimate of turnovers per 100 possessions. The formula is: TOV% = (100 * TOV) / (FGA + 3PA + TOV + 0.39*FTA)

ORB% (Offensive Rebounding Percentage)	An estimate of available offensive rebounds a player grabbed per 100 possessions. The formula is: ORB% = 100 * (ORB * (Tm MP / 5)) / ((Tm MP/5) * (Tm ORB + Opp DRB))
FT/FGA (Free Throws per Field-Goal Attempt)	The free throw factor measures the team's ability to get to the free-throw line and the ability they have to make them. FT/FGA is the formula.

Table 3.3: The Four Factors of Basketball success by Dean Oliver

Additionally to this, The NBA has offered more insightful measurements since the early 2000s that teams and analysts still use today[2]. One of these is the calculation of possessions. They are regarded as the most crucial discovery the NBA made for basketball statistical analysis. Possessions allow coaches to know the number of chances they had to score any number of points. A possession ends when the team scores any number of points or the opponent team regaining possession by grabbing a defensive rebound or a turnover by the team. For possessions to be calculated, manual revision of the play-by-play data counted all the number of free throws that ended a possession, as not all free-throws end a possession. This revision is to find a globalized percentage to better calculate possessions precisely to this female Ecuadorian league. After this review, it was found that 39% of all free throws in the 21 group stage games ended a possession. The possession formula can be found in Table 3.4. With this, possessions per game were calculated, giving a comprehensive insight into every game per team. Thus, points per possession were obtained. The pace is another factor that comes from possessions, as it is the number of overall possessions per the amount of total playing time. Offensive rating predicted the number of points per 100 possessions and was also computed. Furthermore, some other advanced statistics were obtained, such as True Shooting Percentage, which measures the efficiency of all shots taken, the 3 point field goal rate that measures the rate of 3 point shots taken, and the free throw attempt rate measures the rate of 3 point shots taken. These stats give a better insight into the different aspects of the game, just with the box-score data. This data can be observed in Figure 3.3.

More Advanced Statistics	Description
Poss (Possessions)	This is computed by taking into account all the events that end a possession and the previous analysis done before. Poss = FGA + 3PA + TOV - ORB + 0.39*FTA.
PossPG (Possessions per game)	The average number of possessions per game by a team. PossPG = Poss / Games.
PpPoss (Points per Possession)	The average number of points per possession. PpPoss = Pts / Poss.

[2] `https://www.nba.com/thunder/news/stats101.html`[last access: March 29, 2021]

Pace	The average number of possessions per 40 minutes. The formula is Pace = 40 * ((Tm Poss + Opp Poss) / (2 * (Tm MP / 5)))
ORtg (Offensive Rating)	The average number of points per 100 possessions. ORtg = PpPoss * 100.
DRtg (Defensive Rating)	The average number of opponents points per 100 possessions. DRtg = Oppo PpPoss * 100.
NRtg (Net Rating)	The difference between ORtg and DRtg. NRtg = ORtg - DRtg.
FTr (Free Throw rate)	Number of free throw attempts per field goal attempt. Ftr = FTA / (FGA + 3PA).
3PAr (3 Point field goal attempt rate)	The rate of 3 point field goals attempted by total number of field goals. 3PAr = 3PA / (FGA + 3PA).
TS% (True Shooting Efficiency)	A measurement of efficiency regarding all types of shots, field goal attempts from 2 or 3 and the free throw. It can be seen as an extension of eFG%. TS% = PTS / 2*((FGA + 3PA + 0.39*FTA)).

Table 3.4: Advanced Stats used by the NBA applying them to FEB

This procedure is repeated for all the games in the playoff stage of the tournament. It is organized and prepared for usage in the R software. Zuccolotto & Manisera [24] provides an excellent book for basketball data science with plenty of applications in R. They created an R package for free download and usage called BasketballAnalyzeR[3] and is correctly working in R version 3.6 on Ubuntu 18.04. This package requires data to be of the same structure revealed in their book, which is highly similar to the one mentioned above. This package aims for simplicity and flexibility with functions that extend other R plot packages and make them simpler to use. They also have a fourfactors function that allows computing Dean Oliver's four factors of success having the correct data structure.

Now that the data is extracted and organized. Before attempting any statistical analysis, the data will be studied in its nature to know the best methods to be performed on it. Most of the data is numerical. The central limit theorem states that with a large enough quantity of data, even if they are far from being normally distributed, the averages of such data can be shown to be normally distributed [25]. With the low quantity of data available for the female basketball tournament, this theorem does not apply, so the normal distribution of every game variable will be verified using Shapiro-Wilk's test for normality alongside density plots [26]. This is performed for statistical protocol. The plots can be found in Figure 3.4. Shapiro-Wilk's test provides a value. If that value is over 0.05, then the variable can be assumed to be distributed normally [25]. For the most part, most basketball variables are distributed normally, with certain exceptions to the variables 3 point shots made, offensive rebounds, blocks, and personal fouls. This distribution could occur due to the low amount of blocks and 3 point shots made in this female basketball league.

[3] `https://bodai.unibs.it/bdsports/basketballanalyzer/`[last access: April 6, 2021]

gt	tm	poss	ppposs	efg.	tov.	orb.	ft.fga	ts.	X3par	ftr	bp	fbp	X2cp	pntp	ptov
RSG1	VC	89	0.900	0.416	15.4	29.4	0.208	0.455	0.299	0.364	20	15	17	52	37
RSG2	AO	81	0.843	0.518	27.4	63.6	0.158	0.534	0.211	0.298	5	14	7	38	29
RSG3	VO4	90	0.904	0.425	20.7	26.1	0.358	0.502	0.313	0.522	6	33	10	40	35
RSG4	DC	77	0.817	0.492	26.7	30.0	0.067	0.499	0.250	0.133	10	7	9	30	20
RSG5	VC	82	1.034	0.460	17.0	37.5	0.213	0.511	0.360	0.280	36	15	16	34	40
RSG6	AO	72	0.819	0.427	19.5	14.7	0.218	0.476	0.255	0.327	5	14	9	30	14
RSG7	VC	91	0.736	0.295	19.4	30.4	0.329	0.385	0.507	0.493	29	17	19	18	24
RSG8	AO	80	0.676	0.285	20.9	42.1	0.262	0.356	0.231	0.431	15	13	8	22	14
RSG9	UDJ	83	0.637	0.331	25.2	21.6	0.194	0.388	0.355	0.258	24	14	6	24	16
RSG10	PSC	79	0.519	0.242	22.5	21.7	0.177	0.297	0.290	0.290	11	5	6	18	18
RSG11	VC	85	0.814	0.359	22.6	51.0	0.167	0.402	0.513	0.256	20	19	20	24	18
RSG12	AO	80	0.722	0.433	28.0	25.7	0.100	0.451	0.233	0.183	8	6	13	36	28
RSG13	VC	93	0.740	0.407	23.8	36.4	0.107	0.415	0.307	0.280	30	24	9	42	36
RSG14	PSC	86	0.627	0.297	20.2	28.3	0.188	0.341	0.333	0.377	10	16	7	22	10
RSG15	AO	69	0.594	0.294	24.7	21.6	0.216	0.353	0.196	0.353	11	9	2	16	19
RSG16	VC	91	0.760	0.310	13.1	41.4	0.130	0.346	0.283	0.217	12	27	18	44	33
RSG17	AO	71	0.816	0.456	23.4	35.7	0.105	0.467	0.351	0.228	11	14	10	28	30
RSG18	UDJ	74	0.924	0.439	11.0	18.6	0.152	0.468	0.424	0.258	30	8	6	22	23
RSG19	VC	79	0.769	0.459	24.3	20.6	0.082	0.467	0.443	0.180	9	9	3	18	28
RSG20	DC	75	0.839	0.432	19.7	31.4	0.091	0.456	0.394	0.121	7	8	10	32	10
RSG21	PSC	82	0.681	0.327	22.2	17.8	0.364	0.399	0.436	0.709	15	11	7	14	16

Figure 3.3: Advanced Box-Score totals for every winning team of each game of the 2020 Female Basketball regular season phase

The advanced variables are also tested with their density plots and Shapiro-Wilk's test. The plots can be found in Figure 3.5. The following variables are not precisely distributed normally: bench points, fastbreak points, 2nd chance points, points in the paint, points off turnovers, free throw rate, and free throws per field goal attempt. Once again, the amount of data is insufficient to have a normal distribution.

Before continuing to the statistical analysis, this data can be visualized with many different plots. Radial plots and scatter plots will be used by inspiration of Zuccolotto et al. [24]. Radial plots help visualize team and player profiles. The radial plot has numerical values plotted as distances from the center of a circle having many directions dictated by the number of variables [24]. It is important to note that the distance of each point from the center is the focal point for every variable. Also, the area is determined by the data of these variables and can have an overemphasis on high numbers [24]. It is for this reason that the radial plot can be criticized. However, recognizing the game variables to plot can provide an overall team profile with respect to specific game variables. Also, scatter plots are used for displaying values for two variables on a Cartesian graph. These plots can indicate the relationships between the variables and recommend associations of many kinds [24].

3.1.2 Statistical Analysis Phase

Following the nature of the state-of-the-art, performance indicators will be found for this tournament. However, most state-of-the-art only seem to utilize the basic box-score statistics found in Table 3.1. These methods will be replicated, but an additional analysis will also be performed utilizing the advanced box-score stats from Table 3.3 and some from Table 3.4 to reveal more insights for coaches and players alike.

The games are first manually divided into group stage games and elimination games as studies have shown these two stages to differ in performance [3, 14]. Then the games must

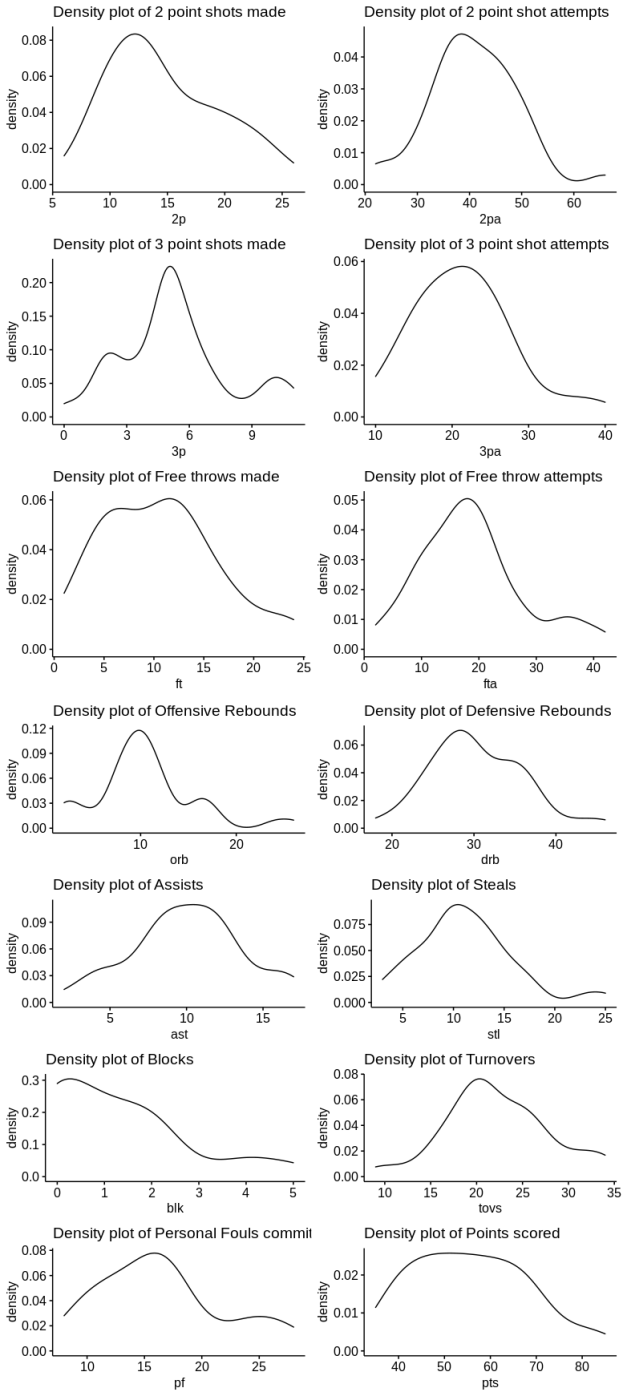

Figure 3.4: Density Plots of every standard box-score stat

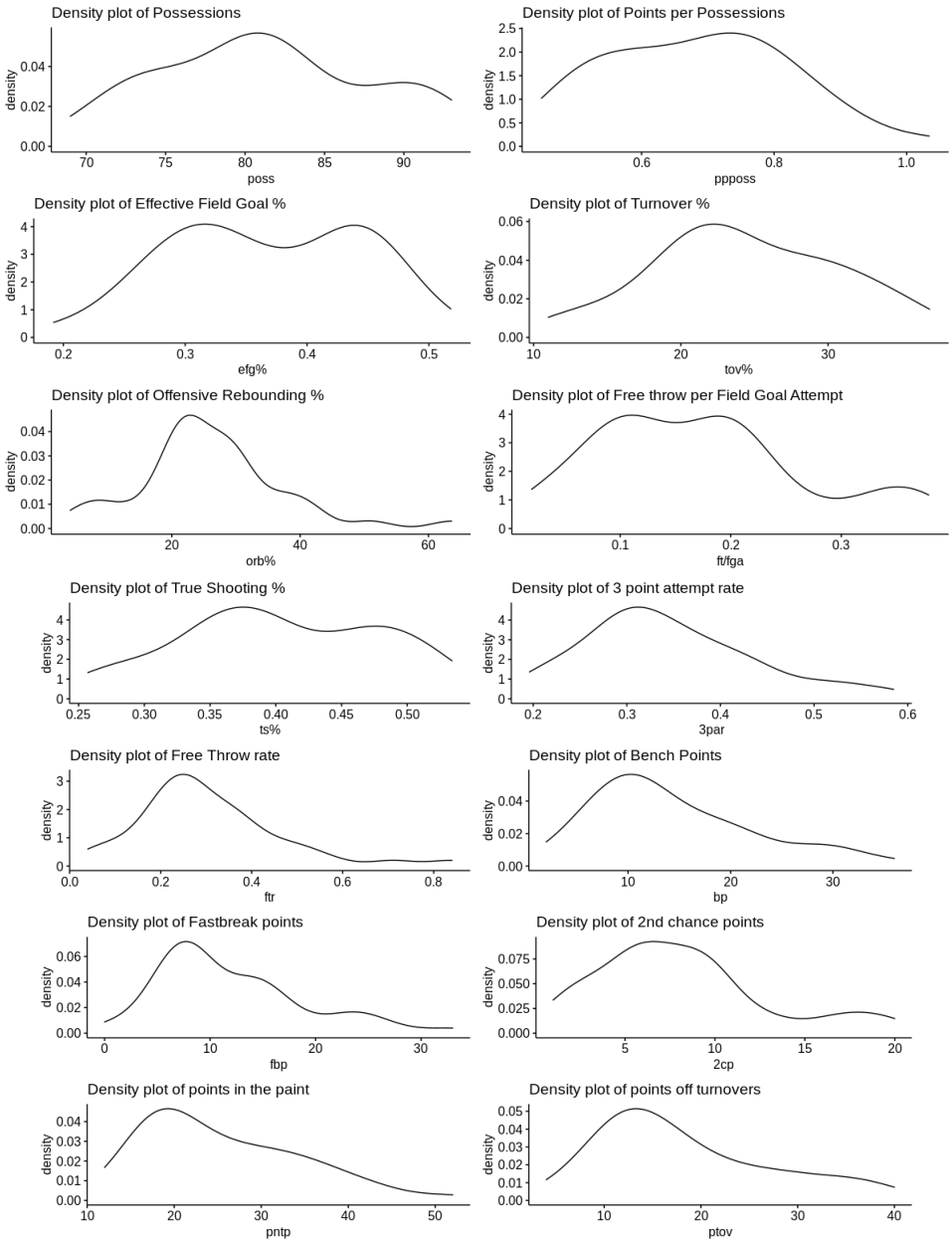

Figure 3.5: Density Plots of every advanced stat

be organized into two groups, balanced games, and unbalanced games. This division will offer a better statistical analysis to identify the variables that differentiate teams that are winning and losing [3, 6, 14, 15]. Cluster analysis is a perfect way to categorize games by simply using the point differentials of every game.

Cluster analysis is a statistical method that groups up individuals of similar characteristics into unknown groups [25]. This analysis is different from classification. There is no pre-defined set of attributes that the observations are to be assigned. The goal of clustering is to discover groups of similar objects and identify interesting patterns in the data [27]. There is no prior information for grouping up the data. Because of this, clustering is known as unsupervised learning, or pattern recognition [27]. This technique is found and can be performed in any field of research. It has been used for classifying different types of depression, classifying domestic roses, classifying fossil remains, etc. [25]. For the grouping of individuals to happen, some measurement of similarity is needed, which are typically defined as distance measurements. This measurement relates to the distance between two observations. This distance will cause the algorithm to place the observation into one group or another [25]. There are various types of cluster techniques, such as hierarchical clustering and non-hierarchical clustering.

One of the most common non-hierarchical clustering methods is the k-means [25]. This algorithm divides the data randomly into K initial clusters specified by the user. The centroids are computed of each initial cluster [25]. Once a new observation is given, the distance measurement is used to find the minimum distance with each centroid and is placed inside the closest cluster. This method is iterative until all observations are assigned to their respective cluster [26]. As mentioned before, K-means will be used to group up games into two categories having the point differential by the distance metric. These clusters will form according to point differential for every game. For each game, the respective points of the winning team are subtracted by the points of the losing team giving the point differential of the match. This procedure is performed for all games, and then it is submitted to the k-means clustering algorithm using the cluster package in R.

At this point, any statistical method can be performed or preferred depending on the purpose. This project will attempt to perform various statistics explaining the different purposes, results, and interpretations. ANOVA is the first method utilized.

ANOVA stands for ANalysis Of VAriance and is commonly used to compare the mean difference of two groups together to find any statistical differences [28]. Performing ANOVA tests a null hypothesis that the means of both groups are equal. In other words, the variable does not have a significant impact on the outcome. In this case, performing ANOVA for every game variable testing winners and losers can mean that the variable is significant for one group or another [28]. The other hypothesis is that the means of both groups are not equal, thus creating a sense that the variable is significant in one form or another. ANOVA yields an F-value, which is the rate of the regression mean by the residual mean [28]. Comparing this F-value to the F-value table can determine its significance or not. ANOVA has various assumptions of the data it is to be performed on. It assumes the data has a normal distribution [26]. There is a homogeneity of variances among groups, and each observation is independent of all the others [26]. Per the last subsection, this basketball data is distributed normally, and a game is independent of every other game. As for homogeneity of variances, this will be assumed as every team can overachieve or underachieve from the average amounts in every game. In basketball statistics, performing

ANOVA can help find the variances of all box-score stats and comparing the values of winning and losing teams to determine the most significant variables and those that are not [3, 6]. So, all box-score stats are tested for ANOVA individually to determine its statistical difference for winning and losing teams.

Obtaining correlation between game variables can also be done utilizing the *basketballAnalyzeR* with a function corranalysis. The BasketballAnalyzeR [24] provides a correlation analysis of variables that will be used to find the correlations of these variables with the current dataset. This function analyzes a pairwise linear correlation among variables, and the term "pairwise" indicates that, even if more than two variables are jointly analyzed with these two functions, leading to the creation of a correlation matrix and its related plot, the linear correlation coefficient remains a measure of bi-variate association, as it evaluates direction and intensity of the linear relationship between all the possible pairs of variables. [24]. The function presents a graphical description of the correlation matrix, which can be performed to locate positive and negative correlations of variables for coaches to interpret and analyze on their strategies.

Once performing these statistical analyses, a supervised approach can be taken using the clusters resulting from the cluster analysis. Now, as ANOVA determined the statistical significance of every box-score statistic individually between teams, the same has to be done, taking all variables into account since all of them occur in a basketball game. Discriminant analysis tackles the issue of describing the groups [25]. There are two objectives regarding the division of groups: one is the description of the group where a linear function of the variables (discriminant function) can elucidate the differences between two or more groups taking all variables into account, and two is regarded with prediction and classification goals were linear or quadratic functions (classification functions) are obtained to classify new observations to one of the already identified groups [26]. At this point, descriptive discriminant analysis is sought after to find the box-score metrics that hold the most weight collectively to compare winning and losing teams [14, 16, 17, 3].

After knowing the variables that most contribute to winning games, creating a decision tree to visualize these variables and their quantities for success can be helpful [20, 5]. CART is an empirical method for creating a binary decision tree [25]. A decision tree presents the data giving rules for every variable, and a general outcome [29]. They are typically used for prediction purposes. In this case, predicting the outcome of a basketball game for a specific team can provide plenty of insights into their performance.

After performing the analysis on the box-score stats, the same analysis will be performed using the advanced stats inspired from NBA analytics and Dean Oliver's four factors of basketball success [9]. These stats can provide more insight for coaches and players using advanced stats.

As mentioned two sections ago, basketball experts Manolo Albán and Paola Herrera in Ecuador were consulted for finding various aspects of the games that can be quantified, yet are not. This is a recollection of the ideas and thoughts that came out of the discussion. The box-score stats do not quantify nearly enough of the defensive skills of the teams and players. Everything is currently too focused on the offensive side of the game. While the offense is essential, it is missing out on the other 50% of the game.

A basketball game is a dance between attempting to score points and not allowing the opposing team to score as many points. With this in mind, specific events can help shed light on the entire basketball game. One event is the shot contest; simply dividing

field goals into contested and uncontested can provide a whole new field into data analysis and bring many insights for coaches regarding their defense and the opponent's defense. Another event found is forced and unforced turnovers, by categorizing turnovers into forced and unforced brings a lot more insights into the value of the defense. Another event is the player or team containment, containing the opponent team or player into an empty possession. Another event is passes allowed in the paint. Having a defense permit passes in the paint can give insight into their defensive strategies and philosophy.

Moreover, finally, dividing the type of rebounds into contested and uncontested rebounds provides a clearer picture. Any player can grab uncontested defensive rebounds. However, players who can grab rebounds beating two or three opponent teams also attempting to grab the ball are much more valuable for player profiling. These events can easily be included in the data.

As this is not included, manual extraction of these events is taken using live recordings of the basketball games from FEB's Facebook page[4]. After obtaining this data and preparing it inside R, the various statistical analysis can be explored once again.

Utilizing these new concepts and giving a logical, numerical value for them can be put to the test for CART to see if they provide even more insight or bring a more precise picture into how a team can attempt to win over their opponents.

[4] `https://facebook.com/FebEcuador/videos/`[last access: March 30, 2021]

Chapter 4

Results and Discussion

In this section, the results and discussions of the various statistical analysis are stated. A critical remark to consider is how sports are not an exact science. So there are many different offensive and defensive strategies that coaches prefer and will be better than other strategies according to the player profile. However, there is no one exact formula for every team. A strategy of one team could be beneficial to them but detrimental to another. Performance is not consistently stable as other teams can adapt. This section aims to present the data and its analysis and discuss the various interpretations that there could be while providing coaches and players with new insights into their strategies and philosophies. There is a big reason every team in the NBA has their own data analytics staff alongside their coaching staff.

4.1 Data Visualization

Using data visualization tools can help coaches and players see the data in a summarized, comparative, and insightful manner. For this first section, plots and graphs are used for coaches and players to read them and provide a more straightforward way to interpret data once taught. Figure 4.1 gives a profile on the shooting variables of every team in the Female Tournament. The plot shows every team has a profile that states how they attempt 2 point field goals above all but only makes a below-average amount. The teams are also not shooting a lot of 3 point field goals, except Victoria Cogarol. However, just like the rest of them, all teams suffer to make 3 point field goals. Regarding free throw shooting, Victoria Cogarol and UDJ attempt an average amount of free throws and make at least half of them, while the other teams attempt a low amount of free throws. This radial plot gives the current state of the Female Basketball tournament, as teams are more likely to generate their points from the 2 point shot.

Another radial plot can be performed with the non-shooting variables found in standard box scores. Figure 4.2 provides an insight into the other variables for team performance. In all teams, blocks were a rare case. All teams have a similar defensive rebounding profile; however, the highlighted teams are UDJ, Deportivo Cuenca, and La Perla SC, who grab more defensive rebounds, followed by Victoria Cogarol and Audaz Octubrino. Offensive rebounding is a more rare occurrence; however, Victoria Cogarol is a team that

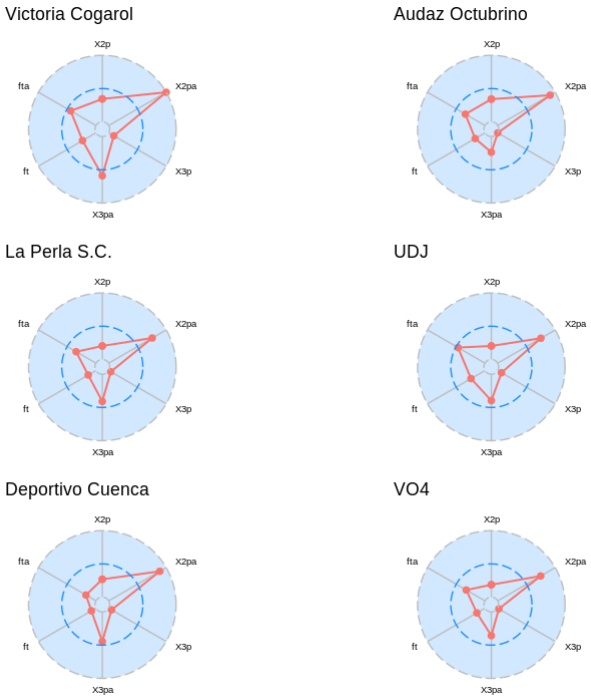

Figure 4.1: Radial Plots for every team regarding shooting statistics

obtains offensive rebounds at a higher amount than the other teams, while other teams like Deportivo Cuenca and VO4 are not far behind. The other remaining teams do not grab as many offensive rebounds. Also, the plot shows how every team commits a high amount of turnovers. Regarding personal fouls, almost all teams commit a decent amount of personal foul. UDJ has a slightly above-average distance of personal fouls, and Audaz Octubrino has a slightly below-average distance of personal fouls. Regarding assists, the teams seem to have a low assist profile. Victoria Cogarol and Audaz Octubrino were able to steal the ball a fair amount more than the other teams. The other teams show a below-average steal profile.

Advanced stats can also be put into graphs to visualize teams' profiles. The WNBA will be compared for conceptual purposes as these numbers are best understood when referencing another basketball league. Figure 4.3 presents a radial graph of some advanced stats that are either percentages or rates, so the graph does not suffer from value bias or inaccuracies. The league average on EFG% is at 0.363. Audaz Octubrino has the highest at 0.418, Victoria Cogarol is next with 0.382, and UDJ being a close second with 0.378 are the teams that average above the league average. This is reflected in the graph as the other teams are close behind UDJ, indicating that the teams do not vary significantly in this variable. Comparing these stats from the 2019 WNBA season, found in the website of basketball-reference[1], the league average for eFG% is 0.473, which is almost 10% less. Going on to the ORB%, the league average is at 26.6%, with Victoria Cogarol having an average of 36%, Deportivo Cuenca at 29.6%, and VO4 with 28.9% being the teams above the league average. The WNBA ORB% league average is at 25.9%, meaning the Ecuadorian league slightly grabs more offensive rebounds than the American League. The

[1] `https://www.basketball-reference.com/wnba/years/2019.html`[last access: April 5, 2021]

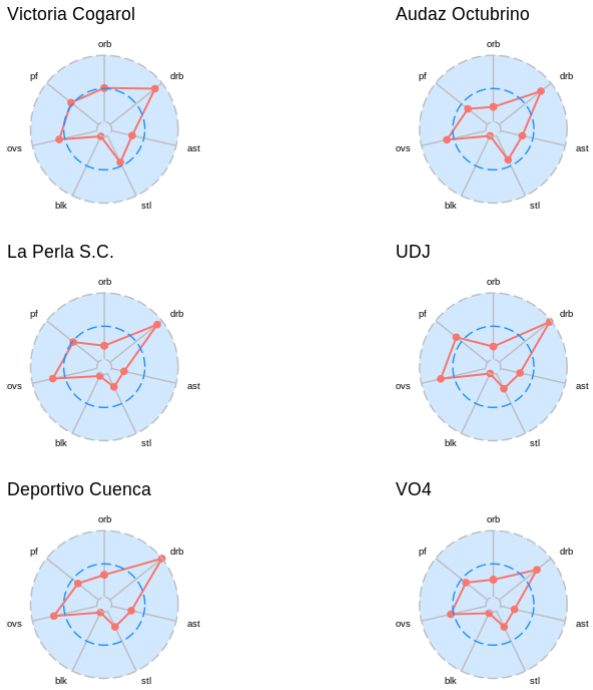

Figure 4.2: Radial Plots for every team regarding nonshooting statistics

plot shows how most teams except Victoria Cogarol tend to commit the same number of turnovers per 100 plays. The league average is 24.3, and Victoria Cogarol is the only team below the average with 19.2. The WNBA has a league average of around 15.1, which can signify the difference in ball-handling between the two leagues. Continuing to the last factor of success by Dean Oliver [9], FT/FGA, the league average is 0.169, and UDJ has the highest at 0.245, Audaz Octubrino has 0.194, followed by Victoria Cogarol with 0.177. This average is not far off from the WNBA that has an average of 0.2. The plot shows the minimum distance of all the other variables. The following two advanced metrics are the rate of free throws that the team takes and the rate of 3 pointers attempted. Most teams have a relatively similar rate for both these shot attempts. The league average for both is 0.306 and 0.345, respectively. Audaz Octubrino has a higher rate of free throws than three-pointers, while Deportivo Cuenca has a higher rate for 3 point shots than free throws. Finally, for True Shooting Percentage, Audaz Octubrino is right on the edge of the circle, signaling this as the highest value at 0.457. The league average is 0.4, which is a lot less than the WNBA average at 0.516.

Overall these provide insights into their tendencies, and coaches can develop strategies accordingly. For example, looking at Audaz Octubrino's advanced stats as a whole, they have a low rate of three-point attempt rate, do not generate and make many free throws, turn the ball over a fair amount, do not grab as many offensive boards, and is the most efficient shooting team.

Taking advantage of some advanced statistics that the FEB does track. A radial plot can give insight into the different events when points are scored. Figure 4.4 provides this for the variables explained in section 3. Victoria Cogarol has the highest profile for all variables than the other teams. They get lots of points in the paint, points off opponent's

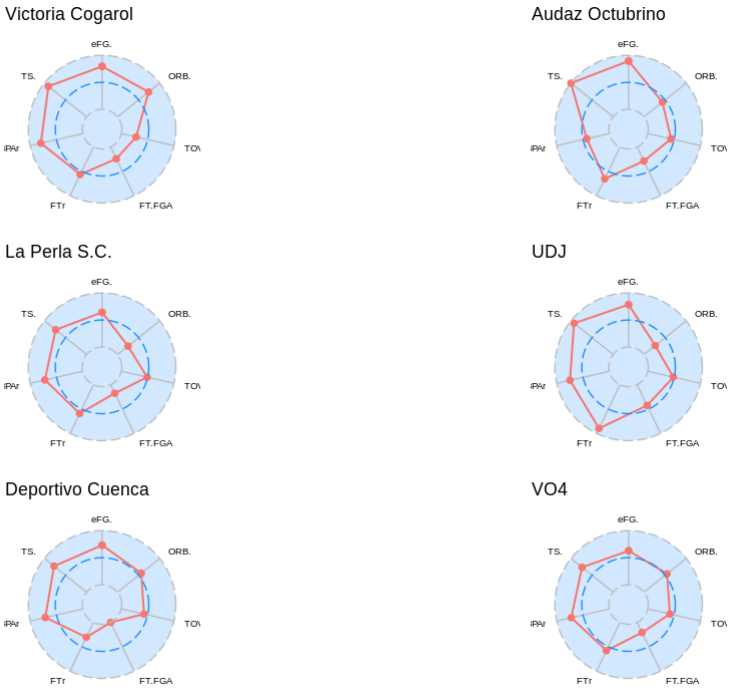

Figure 4.3: Radial Plots for Advanced stats of every team

turnovers, fastbreak points, and 2nd chance points. Audaz Octubrino also has a high point in the paint profile alongside points off turnovers. Octubrino also has a pretty low bench point, fastbreak, and 2nd chance point profile. Overall, for the rest of the teams, points in the paint seem to be a high profile for all. This female tournament has teams making a significant amount of points in the paint. Not many 2nd chance points, bench points, and fastbreak points can be registered compared to Cogarol. Thus, this league can be seen to consist of teams that create more points in the paint and off turnovers than bench points, fastbreak points, and 2nd chance points.

Scatter plots can also compare two variables together and determine relationships or discriminate teams from each other. The following Figure 4.5 is a scatter plot of offensive rating and defensive rating for each team, highlighting the team with a color that is their net rating, which is the difference between them. Victoria Cogarol and Audaz Octubrino are the two top teams in the initial phase, but the stats show why they are on top. Victoria Cogarol is exceptionally ahead of the other teams in terms of offensive rating and defensive rating. Between the colors, yellow and red are positive net ratings, while yellow and blue are negative net ratings. VO4 was the worst team to perform in this tournament, and they had nothing going on their side in terms of these stats.

The next scatter plot attempts to plot pace and points per possession while using True Shooting percentage as the color to provide more insight into performance. Figure 4.6 displays this, and there is a very noticeable observation. At the far end of the plot, Victoria Cogarol stands on top with the highest pace of all the teams with 86.14, and going to the other side of the plot, Audaz Octubrino has the slowest pace with 75.86. Victoria Cogarol is averaging about 11 more possessions per game than the other top team in the tournament. Cogarol has a higher point (0.82) per possession count than Octubrino (0.74). However,

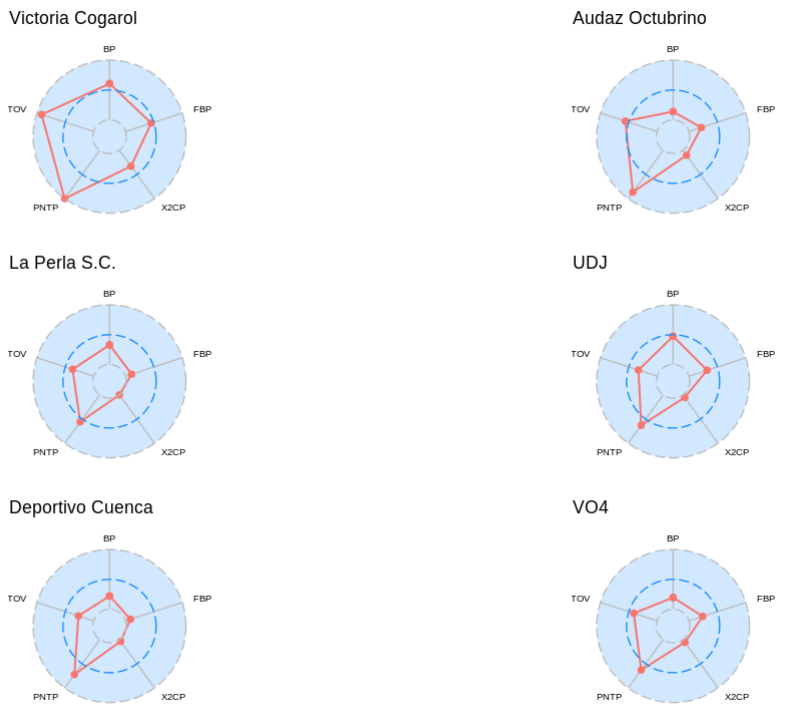

Figure 4.4: Radial Plots for Different events for points scored of every team

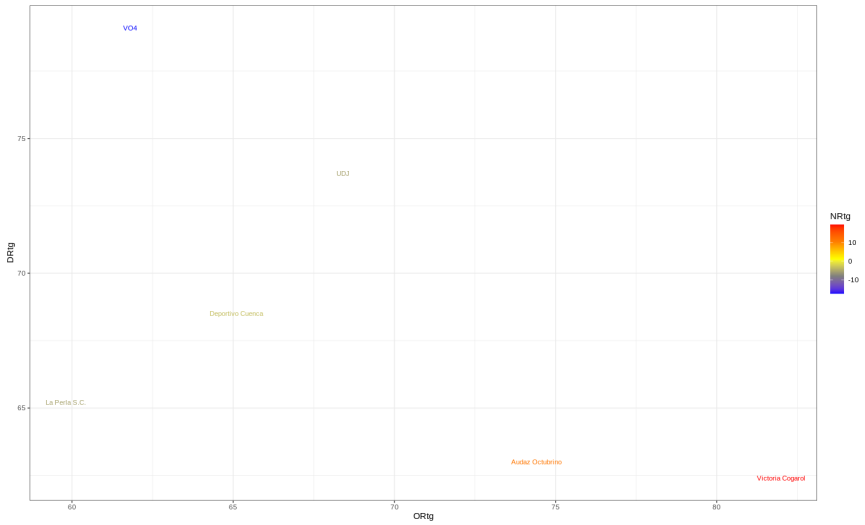

Figure 4.5: Scatter plot: Offensive vs Defensive Rating plus Net rating

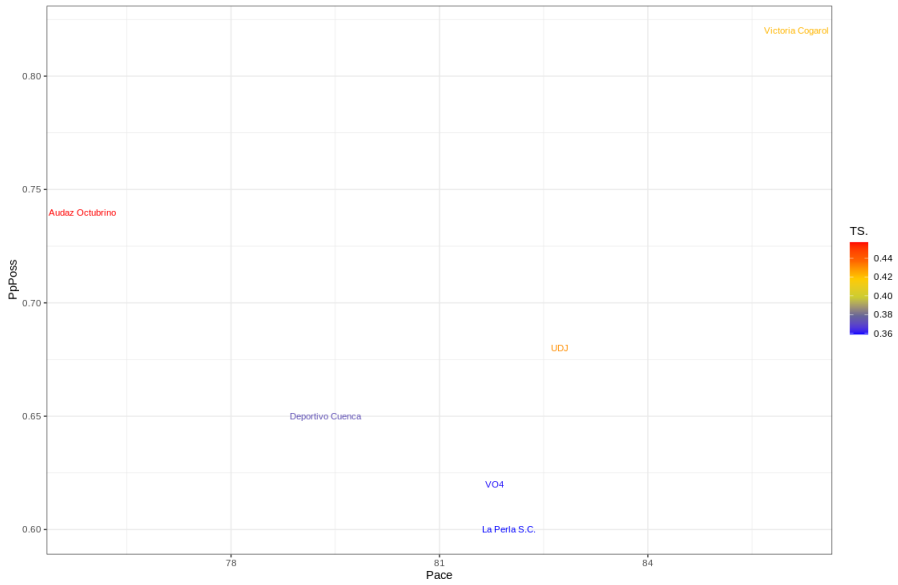

Figure 4.6: Scatter plot: Pace vs Points per possession plus True Shooting Efficiency

the latter has a higher true shooting efficiency. For perspective, the WNBA average pace was 77.7 while this league average was 81.31, which states that these women play at a faster rhythm than the American League.

Once again, these data visualization tools can give coaches and players meaningful data into their performance. When coaches attempt to hold their players accountable, the numbers do not lie, and they can backup for making decisions, such as benching a player who is playing a lot and contributing little to their role.

4.2 ANOVA

The following section shows the ANOVA test between the same game variables of the different teams in a game, winner vs. loser, to determine any distinguishable variables. Table 4.1 places the results of each game variable, from standard box-score to advanced stats. The analysis yields that the 2 point shot is significant in differentiating winning teams from losing teams. Winning teams also grab more offensive rebounds than their opponents. Winning teams also seem to pass the ball that results in more points than their opponents, and they also steal the ball more, forcing more turnovers over the other team. Turnovers are also a factor. Going towards the advanced stats, teams that have generated more points per possession often win. As offensive rebounds and turnovers are indicators, so are offensive rebounding percentage and turnover percentage. A team is also winning when they have a higher true shooting percentage. As for the variables concerning points, all except bench points state that teams are winning when they score more.

Game Variable	ANOVA results (p-value)
2 point shots made (X2p)	0.0036*
2 point shots attempted (X2pa)	0.0206*

3 point shots made (X3p)	0.422
3 point shots attempted (X3pa)	0.555
Free throws made	0.112
Free throws attempted	0.229
Offensive Rebounds	0.0289*
Defensive Rebounds	0.269
Assists	0.0304*
Steals	0.00994*
Blocks	0.46
Personal Fouls	0.598
Turnovers	0.00355*
Possessions	0.945
Points per Possession	0.000025*
Effective Field Goal Percentage	0.139
Offensive Rebounding Percentage	0.0065*
Turnover percentage	0.00087*
Free throw per Field Goal Attempt	0.447
True Shooting Percentage	0.0475*
3 point attempt rate	0.378
Free Throw Rate	0.065
Bench Points	0.47
Fastbreak Points	0.046*
2nd chance points	0.0048*
Points in the Paint	0.0333*
Points off turnovers	0.0000346*

Table 4.1: Anova Results for standard and advanced box-score statistics

4.3 Correlation Analysis

The number of points scored at the end of the games depends on all the events on the court. Figure 4.7 demonstrates the correlation analysis taking box-score data of the winning teams for all 21 games played. There are two parts to the plot. The left section is the graphical representation of the correlation matrix. The diagonals simply state the variable name, while the lower-left triangle is the values of the Pearson correlation coefficients. Any entry of the matrix is marked with an X if it is not statistically different from zero [24]. The upper right triangle provides stylized ellipses to give a clue to the relationship of every coefficient. It is also color-scaled, as red denotes a positive correlation, blue denotes a negative correlation. The right part of the plot is the correlation network with the same color scale as on the left. This takes the significant correlation coefficients and joins them by a certain intensity of red or blue depending on its value [24]. Now examining the Figure 4.7, for teams that won their games, there are no negative correlations which can mean that the various events and actions of the games were not negatively impacting other events and actions. The apparent correlations are between the different types of

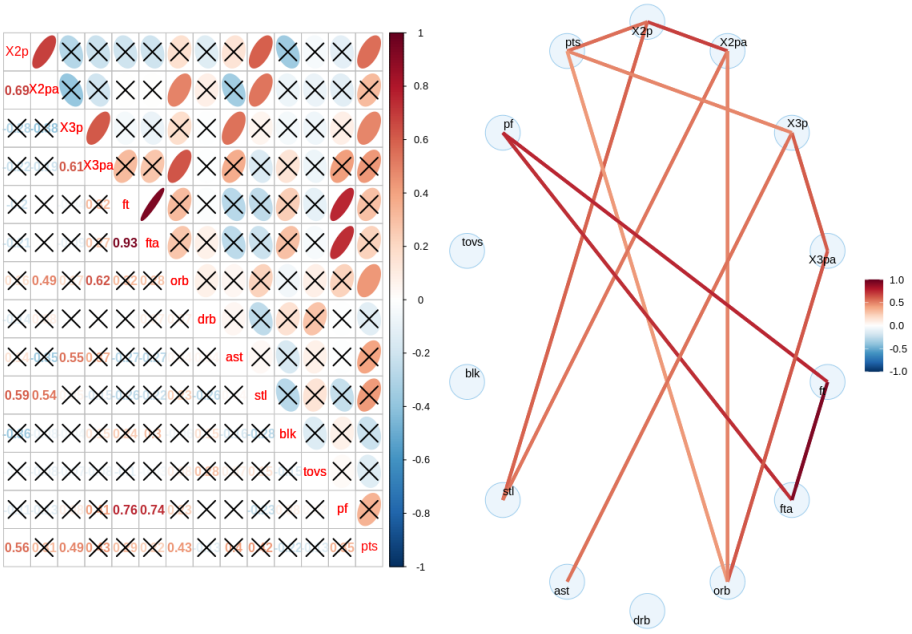

Figure 4.7: Correlation Analysis Plot of standard box-scores for the winning teams of every game.

shot attempts and makes. There is a strong correlation between 2point shots made (X2p) and 2 point shot attempts (X2pa), with three-pointers and free throws. Also, there is a strong correlation between personal fouls (pf) and free throws made (ft) and attempted (fta). Offensive rebounds (orb) are positively correlated with 2, and 3 point shot attempts and points. Assists are positively correlated with 3points made. Steals are correlated with 2 point shot attempts and makes. Winning teams are likely to have more 2 point shots made and attempted while stealing the ball more. Points are correlated with 2 and 3 point shots made and offensive rebounds. So they also get their points from 2 and 3 point shots and get more opportunities at those shots with offensive rebounds. A vital remark to make is how defensive rebounds (drb), turnovers (tovs), and blocks (blk) are not correlated to any other variable. Blocks were pretty rare compared to the other variables. This result means that winning teams turning the ball over has no impact on the other aspects of the game, and the same can be said about defensive rebounds. These results are merely what this dataset is showing, another basketball dataset can show an entirely different graph, and different interpretations would be made.

Going on to the loser side of things, Figure 4.8 provides a correlation analysis of game variables. In this instance, there are negative correlations between several variables. 3 point shots are negatively correlated with 2 point shots made and attempted. Losing teams depict a tendency to either attempt and make a lot of 2 point shots while failing to convert their 3 point shots. 3 point shot attempts are also negatively correlated with 2 point shots made. These teams would stop taking three-point shots the more 2 point shots they would make, There is also a negative correlation of turnovers with 3 point shot attempts. Teams who lose the ball more shoot fewer 3 point shots. As in the previous plot, there are positive correlations between the same shots attempted and made. Offensive

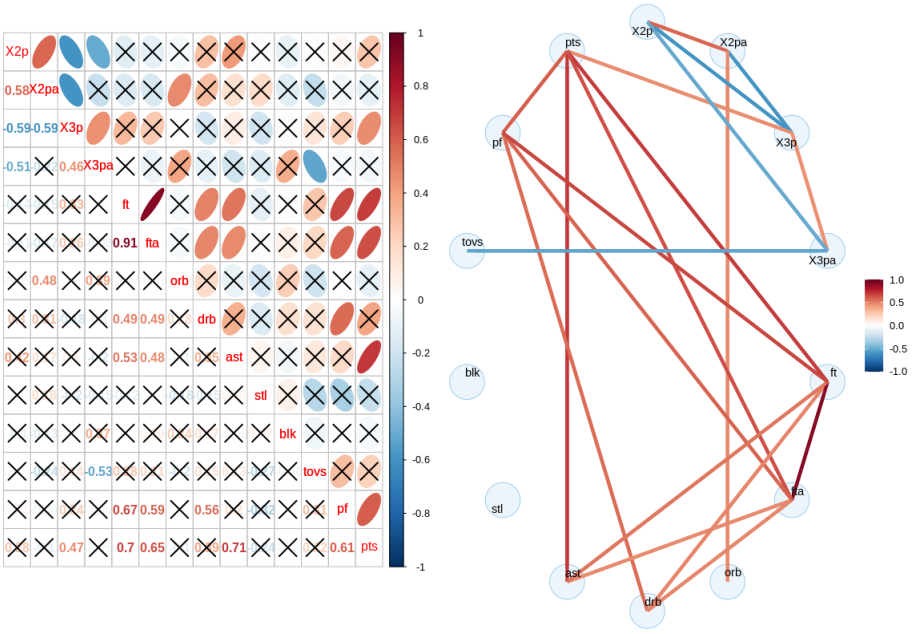

Figure 4.8: Correlation Analysis Plot of standard box-scores for the losing teams of every game.

rebounds are also correlated to 2 point shot, while defensive rebounds are correlated with fouls and getting to the free-throw line. Assists are unusually correlated with free throws. Points are positively correlated with personal fouls, free throws made and attempted, and 3 point shots. The losing teams are not getting statistically significant points from the 2 point line. They do not convert enough 2 point shots to be on the winning side of the game. Finally, blocks and steals are not correlated to any other variable.

After finding the correlation of all standard box-score stats between each other, the advanced stats will be used for finding their correlations. Different variables can be submitted to correlation analysis for interpretation and performance purposes. In this case, the advanced box-score stats are tested with each other alongside points from the standard box-score stat. Figure 4.9 depicts the winning teams. There is a solid correlation between free throw rate (ftr) and free throws per field goal attempts (fg.fta). Another strong correlation is found between effective field goal percentage (efg.) and true shooting percentage (ts.). Winning teams that have a high efg will also have a high ts. These two variables are also strongly correlated to points per possession (ppposs). Being more efficient in shooting variables leads to more points and thus more points per possession. There is a negative correlation between free throws per field goal attempt and effective field goal percentage. It seems that the winning teams are more effective from the 2 and 3 point shots than they are attempting and making free throw shots. Points are positively correlated to possessions, points per possessions, effective field goal, and true shooting percentages. To sum it up, winning teams generate more points per possession and shoot at a higher efficiency than their opponents.

There seem to be no negative correlations between the variables from the losing teams. As seen above, effective field goal percentage is strongly correlated with true shooting

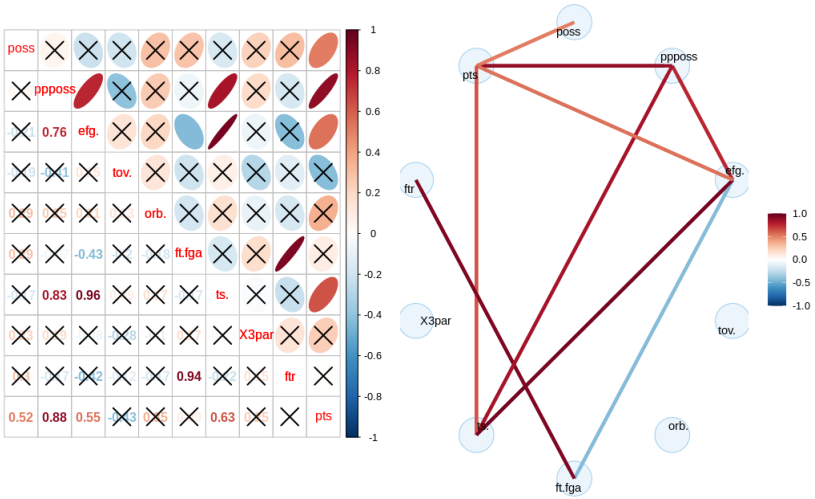

Figure 4.9: Correlation Analysis Plot of advanced stats for the winning teams of every game.

percentage, as are these two variables with points per possession, and as is free throw rate and free throw per field goal attempt. Turnover rate is correlated with an effective field goal percentage and true shooting percentage. Both free-throw variables are correlated to points per possession and true shooting percentage. The losing teams are converting points through their free throws, which increases their true shooting. While free throw per field goal attempt has a positive correlation with possessions, the more free throws they take, the more possessions they obtain. Points are correlated to free throw rate, true shooting percentage, effective field goal percentage, possessions, points per possession, and free throw per field goal attempts. The losing teams seem to find more points through free throws than the winning teams. Both plots show that offensive rebounding rate and 3 point shot rate are not correlated to any other variable.

4.4 Discriminant Analysis

The discriminant analysis results are found in Table 4.2. The means of each variable per class are shown and their linear discriminant coefficient. This calculation is done to find the differences in these variables between the winning and losing teams. There is a reclassification of 85.7%. The variables that discriminate winners from this analysis seem to be 2, and 3 point shots made, offensive and defensive rebounds, and steals. In this league, the winning teams are making more 2 and 3 point shots, grabbing more rebounds, and stealing the ball from their opponents more. Prediction can also be performed using the discriminant functions[25]. For this, the playoff games with the same standard box score are used. Even though there were only 7 games, it seems it has an accuracy of 85.7%. The two games that misclassified were very close games where the team that won did so by a couple of points.

Discriminant analysis is done with the advanced stats and is in Table 4.3. There is a reclassification percentage of 85.7%. The game variables that discriminate the winners

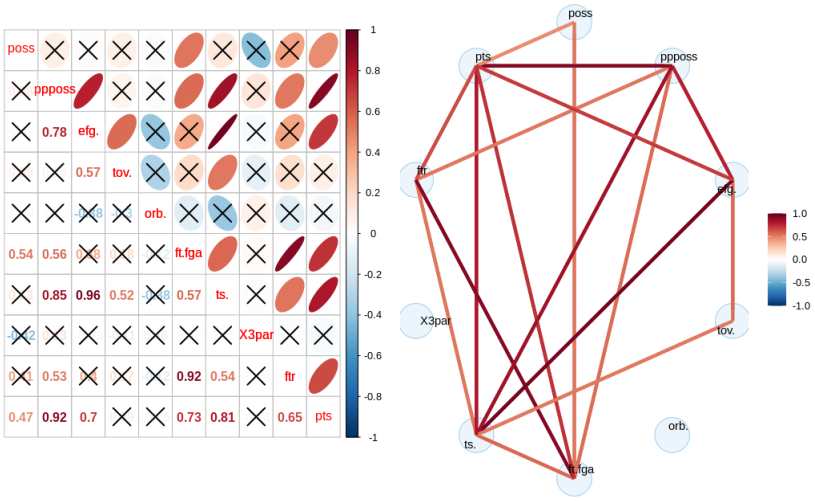

Figure 4.10: Correlation Analysis Plot of advanced stats for the losing teams of every game.

Game Variable	Means Winning teams	Means Losing teams	LD Coefficients
2 point shots made (X2p)	16.857	12.619	-0.17941396
2 point shots attempted (X2pa)	43.71429	37.7649	0.11197222
3 point shots made (X3p)	5.619048	4.95238 1	-0.14066054
3 point shots attempted (X3pa)	22.14286	20.90476	0.09996019
Free throws made	12.142857	9.238095	0.21696319
Free throws attempted	20.42857	17.00000	0.07327028
Offensive Rebounds	12.238095	8.809524	-0.16507673
Defensive Rebounds	31.33333	29.33333	-0.08190490
Assists	11.238095	8.857143	0.02846764
Steals	13	9.333333	-0.07722127
Blocks	1.142857	1.476190	0.04320103
Personal Fouls	16.04762	16.95238	0.08410657
Turnovers	19.80952	24.76190	0.15500896

Table 4.2: Discriminant Analysis for standard box-score statistics of winning and losing teams

Game Variable	Means Winning teams	Means Losing teams	LD Coefficients
Possessions	81.38095	81.23810	0.017317508
Points per Possession	0.7700476	0.6025238	-12.742132403
Effective Field Goal Percentage	0.3859048	0.3492381	46.311545436
Offensive Rebounding Percentage	30.74286	21.42857	-0.005492896
Turnover percentage	21.31905	27.59048	-0.005492896
Free throw per Field Goal Attempt	0.1850476	0.162	15.924347801
True Shooting Percentage	0.4270476	0.3802381	-38.120430484
3 point attempt rate	0.3325714	0.3580476	4.944606561
Free Throw Rate	0.3122857	0.2995238	-4.404851985

Table 4.3: Discriminant Analysis for advanced statistics of winning and losing teams

from the losers are the following variables: points per possession, turnover and offensive rebound percentage, true shooting percentage, and free throw rate. Teams score more points per possession than their opponents and have a better true shooting percentage. Effective field goal percentage is once again not discriminant, as shown in ANOVA. This result is an interesting insight as both measure shooting efficiency, and it shows which of them has more impact on winning. The turnover percentage seems to have a small, almost insignificant impact, while in Table 4.2, turnovers were not a factor for winning teams, but instead, losing teams committed more turnovers. The accuracy of the discriminants was determined by using the advanced playoff stats. There is a 71.4% accuracy of the discriminants. Of course, playoff games are different from regular-season games and tend to be more competitive. However, performance from the regular season can indicate patterns for playoff season [3, 7, 1].

4.5 CART

CART can be handy for visualizing the outcomes of teams and for future predictions. They are also very versatile as CARTs can be computed for different variables outcomes. Figure 4.11 is the tree for wins, and it tells us that teams who commit over 26 turnovers will lose their game. A team is also likely to lose if they commit less than 26 turnovers, less than 10 assists, and make less than 10 free throws. This tree states that teams will win, committing less than 26 turnovers, making more than 10 assists, and converting any number of 2 point shots. Figure 4.12 depicts a CART for points scored. In this case, teams score the highest amount of points at 72 73 when they have more than 10 assists, block

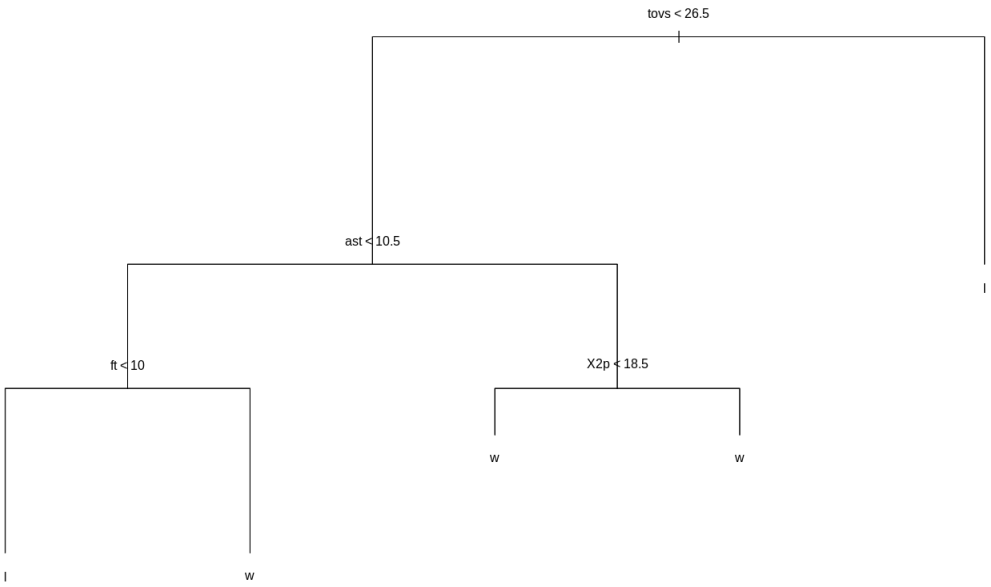

Figure 4.11: CART for wins using standard box-score stats

the ball at least once, and attempt more than 20 3 point shots. The lowest amount of points scored by a team happens when they have less than 10 assists, less than 14 offensive rebounds, and have no blocks.

The CART can also be done with advanced stats to determine more clues about winning and losing teams. Figure 4.13 is similar to the wins CART of standard box-scores. Teams having a high turnover percentage will lose the game. A team is winning if they convert over 0.7295 points per possession, basically scoring more than their opponent. If they go under this amount, they can still win if 0.17 free throws are attempted per field goal attempt.

A CART can also be done for determining the number of points a team can score with these advanced stats. Figure 4.14 seems to be highly focused on points per possession, which is pretty self-explanatory. The more points per possession a team averages, the more points they will score if there are many possessions. So as this graph does not tell us much, another CART can be performed, omitting these variables to obtain another result. Figure 4.15 presents a more insightful vision. Teams with a high true shooting percentage, over 40%, will score more points when they shoot more than 10% of their field goals in free throws and have a turnover percentage lower than 21.55%. If a team has a higher turnover percentage, they will score about 9 points less.A team shooting under 40% of their true shooting percentage can score a decent amount of points if they go for more offensive rebounds, having higher than a 28% offensive rebounding percentage. A team will score the least amount of points if they are shooting under 37% for their true shooting percentage.

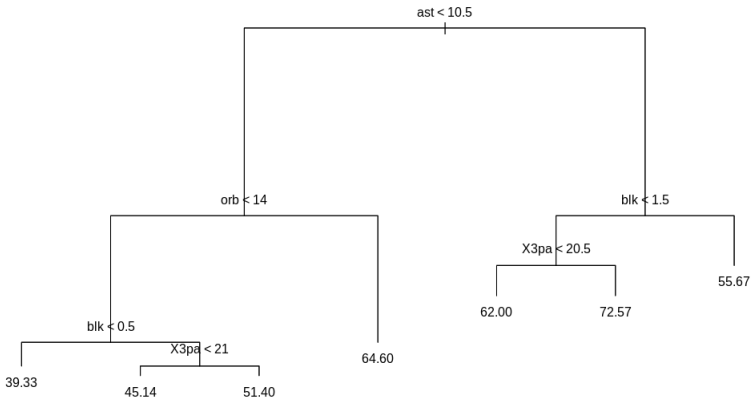

Figure 4.12: CART for points using standard box-score stats

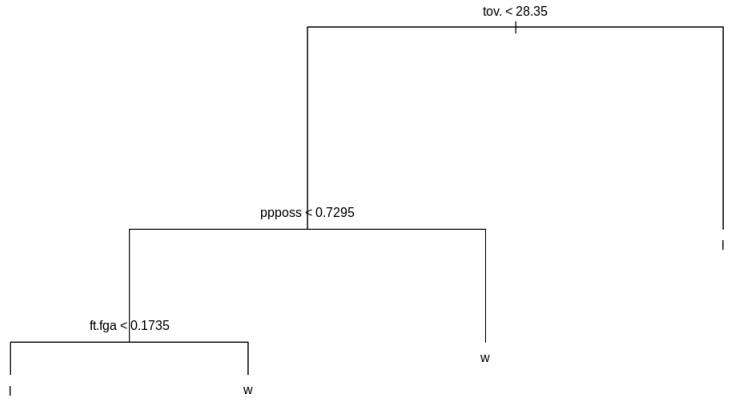

Figure 4.13: CART for wins using all advanced stats

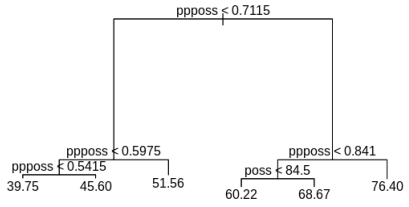

Figure 4.14: CART for points using all advanced stats

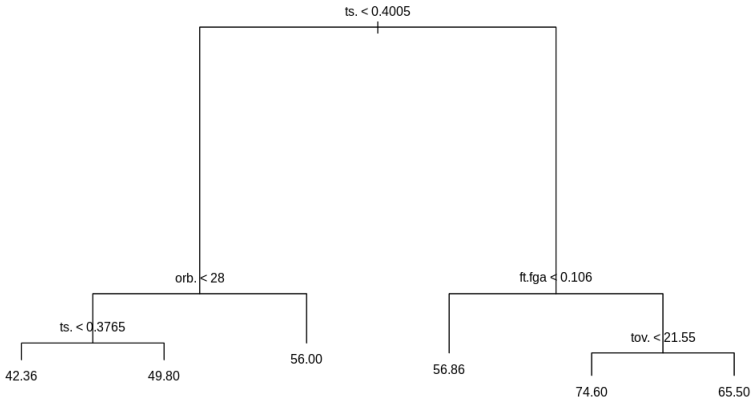

Figure 4.15: CART for points using advanced stats excluding possessions and points per possessions

Chapter 5

Conclusions and Future Work

This study introduces data analysis to sports that contain many variables that explain and lack many variables to explain it entirely. Data were extracted from the female tournament of Ecuadorian Basketball from 2020-2021. This data was worked through various data visualization plots, such as radial plots and scatter plots, to statistical methods as cluster analysis, ANOVA, correlation analysis, discriminant analysis, and CART. All of this is done to provide insights into the data for coaches, players, and fans alike. In various methods, the teams in this country data of basketball obtain most of their points from the 2 point shot. An important remark is that using these statistical methods on another dataset, basketball games from other countries or international competitions will yield different results. That is the beauty of this. Sports do not have a clear formula for every team and player. Sports teams consist of humans, and the abilities of humans will always vary. However, utilizing data and big data analytics can help sports and sports teams to learn more and find more patterns into the sport to create and attempt solutions that can change the game. The best example of this is the NBA that is years ahead in the basketball world.

After talking with two basketball experts of the country, there are cultural issues with the sport. Players and coaches are highly outdated in their thoughts and are unwilling to change. This is an issue with many fields and areas in the country where Yachay Tech was born, to change the country through science and innovation—creating a country with high scientific production and exportation, rather than importing everything. Sports is just another aspect that Ecuador is behind on that science and technology can help revolutionize. Future works can include the various amount of things and ideas. Using the same data with different and more advanced statistical methods such as Markov chaining and Bayesian theory can be applied to provide profiles on players and teams regarding the offensive and defensive aspects separately. Data analysis can be shifted towards working with a specific team and its players, providing insights into the different line-ups the team's rotational players can have and provide results about the best team line-ups while providing the variables that make them the best team can offer. Proposing a national team using player data from all the teams can also be done by extracting each player's strengths and weaknesses and conforming to a logical team that attempts to touch on all aspects of the sport. The same amount of data can be extracted from other countries, such as Latin America, to compare the different leagues, tendencies, and patterns.

Also, as more data becomes available, there is so much specific analysis that can be made. For example, if the same video system from the NBA comes into the country, there is an infinite amount of analysis and fields that can come into play. Information can be extracted regarding basketball plays and finding the efficiency of specific offensive basketball plays to specific defensive basketball strategies used against them. Extracting player movements on the court can also become a significant resource for coaches, personal trainers, and physical therapists. For example, a player had a bad shooting game, and the player movement data detects that the angle of the player's jump shot was too low than her average jump shot release angle. This could be due to fatigue, or the upper or lower muscles inhibiting the jump shot. Coaches and personal trainers can then get to work with this information giving their player more rest and attacking the parts of her body that can be causing this issue. This information can also be helpful for coaches and their decision-making. They were finding the effectiveness of their substitutions and timeouts regarding the timing. A coach could be substituting players too early or too late to positively impact or are calling timeouts at incorrect moments of the game. Regardless of the data available, much can still be done, and it is up to the person's creativity as to what to do with this data. This study attempts to show this.

Bibliography

[1] B. Alamar, *Sports Analytics: A guide for coaches, managers, and other decision makers.* New York: Columbia University Press, 2013.

[2] Z. Ivanković, M. Racković, B. Markoski, D. Radosav, and M. Ivković, "Appliance of neural networks in basketball scouting," *Acta Polytechnica Hungarica*, vol. 7, no. 4, pp. 167–180, 2010.

[3] J. García, S. Ibáñez, R. D. Santos, N. Leite, and J. Sampaio, "Identifying basketball performance indicators in regular season and playoff games," *Journal of Human Kinetics*, vol. 36, pp. 161–168, 2013.

[4] V. D. S. C. Tian, M. Caine, and S. Swanson, "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data," *Applied Sciences*, vol. 10, no. 1, 2020.

[5] P. Zuccolotto, M. Manisera, and M. Sandri, "Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions," *International journal of sports science coaching*, vol. 0, no. 0, pp. 1–21, 2018.

[6] D. de Rose, "Statistical analysis of basketball performance indicators according to home/away games and winning and losing teams," *Journal of Human Movement Studies*, vol. 47, pp. 327–336, 2004.

[7] J. Sampaio, M. Janeira, S. Ibanez, and A. Lorenzo, "Discriminant analysis of game-related statistics between basketball guards. forwards and centres in three professional leagues," *European Journal of Sports Science*, vol. 6, no. 3, pp. 173–178, 2006.

[8] J. Demenius and K. Rasa, "The benefits of advanced data analytics in basketball: Approach of managers and coaches of lithuanian basketball league teams," *Baltic Journal of Sport and Health Sciences*, vol. 104, no. 1, pp. 8–13, 2017.

[9] D. Oliver, *Basketball on paper: Rules and Tools for Performance Analysis*, Washington D.C., 2003.

[10] R. Shah and R. Romijnders, "Applying deep learning to basketball trajectories," 2016.

[11] A. Maymin, P. Maymin, and E. Shen, "Nba chemistry: Positive and negative synergies in basketball." *International Journal of Computer Science in Sport*, vol. 12, pp. 4–23, 2013.

[12] A. Álvarez, E. Ortega, M. Gómez, and J. Salado, "Study of the defensive performance indicators in peak performance basketball," *Revista de Psicología del Deporte*, vol. 18, no. 3, pp. 379–384, 2009.

[13] L. Yongjun, L. Wang, and F. Li, "A data-driven prediction approach for sports team performance and its application to national basketball association," *Omega*, vol. 98, 2019.

[14] J. Sampaio and M. Janeira, "Statistical analyses of basketball team performance: understanding teams' wins and losses according to a different index of ball possessions," *International Journal of Performance Analysis in Sport*, vol. 3, no. 1, pp. 40–49, 2003.

[15] G. Csataljay, P. O'Donoghue, M. Hughes, and H. Dancs, "Performance indicators that distinguish winning and losing teams in basketball," *International Journal of Performance Analysis of Sport*, vol. 9, pp. 69–66, 2009.

[16] J. Sampaio, S. Ibanez, M. Gomez, A. Lorenzo, and E. Ortega, "Game location influences basketball players' performance across playing positions," *International Journal of Sport Psychology*, vol. 39, no. 3, pp. 205–216, 2013.

[17] S. Ibanez, J. Sampaio, S. Feu, and A. Lorenzo, "Basketball game-related statistics that discriminate between teams season-long success," *European Journal of Sport Science*, vol. 8, no. 6, pp. 369–372, 2008.

[18] M. Huang and Y. Lin, "Regression tree model for predicting game scores for the golden state warriors in the national basketball association," *Symmetry*, vol. 12, no. 5, pp. 835–856, 2020.

[19] L. Grassetti, R. Bellio, L. D. Gaspero, G. Fonseca, and P. Vidoni, "An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data," *IMA Journal of Management Mathematics*, vol. 0, no. 0, pp. 1–25, 2020.

[20] M. Migliorati, "Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms," *Electronic Journal of Applied Statistical Analysis*, vol. 13, no. 2, pp. 454–473, 2020.

[21] M. Sandri, P. Zuccolotto, and M. Manisera, "Markov switching modelling of shooting performance variability and teammate interactions in basketball," *Journal of the Royal Statistical Society*, 2020.

[22] R. Metulini, M. Manisera, and P. Zuccolotto, "Space-time analysis of movements in basketball using sensor data," *Firenze University Press*, 2017.

[23] ——, "Modelling the dynamic pattern of surface area in basketball and its effects on team performance," *Journal of Quantitative Analysis in Sports*, vol. 14, no. 3, pp. 117–130, 2018.

[24] P. Zuccolotto and M. Manisera, *Basketball Data Science With Applications in R.* Boca Raton, FL: Taylor Francis Group, 2020.

[25] A.Afifi, S. May, R. Donatello, and V. Clark, *Practical Multivariate Analysis.* Boca Raton, FL: CRC Press, 2020.

[26] A. Rencher, *Methods of Multivariate Analysis Second Edition.* New York, NY: John Wiley Sons, Inc, 2002.

[27] D. Gunopulos, "Clustering overview and applications," *Encyclopedia of Database Systems*, pp. 383—-387, 2009.

[28] B. Tabachnick and L. Fidell, *Experimental designs using ANOVA.* Thomson/Brooks/Cole, 2007.

[29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning With Applications in R.* Springer, 2013.