Men's U-Sports Basketball Analysis

Michael Armanious

2019-09-12

Contents

Chapter 1

Datasets

The data that has been collected comes from two sources: the Synergy database [2] and the OUA website [3]. It is game by game data for every team in the U Sports division in the Ontario University Athletics conference during the regular seasons from 2015/2016 to 2018/2019. The datasets extracted from Synergy contain insufficient data from the 2014/2015 season so only the data from 2015/2016 to 2018/2019 will be used from the data collected from Synergy. Six datasets were extracted per game and aggregated. These datasets are Play Types, Sets, Shots, Transitions, General Statistics, and Player Statistics. The data scraping functions that were used can be found in the appendix.

1.1 Play Types

This dataset shows the number of different types of plays per game.

Table 1.1: Features & Descriptions of Play Types Dataset (Source: Synergy Database)

Features	Description
ID	a unique ID to reference the games (row data)
Team	Name of the team
Season	the year of the regular season that the game took place
All Isolation	Number of possessions that have an isolation play
All Offensive Rebounds	Number of possessions that involve an offensive rebound
All P.R Ball Handler (BH)	Number of possessions involving a pick & roll ball handler
All Possessions	Number of possessions in a game

Features	Description
All Post-Up	Number of possessions involving a post-up play
Cuts	Number of possessions involving cuts
Handoffs	Number of possessions involving a handoff play
Isolation Defense Commits	A category of an isolation play where the defense commits
Isolation Single Covered	A category of an isolation play where no defense commits
Miscellaneous Possessions	Number of possessions that do not fit a certain category
Off Screens	Number of possessions involving an Off Screen
Offensive Rebound PutBack	A category of Offensive Rebounds where ball gets tipped in
Off Reb Reset Offense	A category of Offensive Rebounds where the offense resets
PR BH Defense Commits	A category of pick & roll BH where defense commit
PR BH Single Covered	A category of pick & roll BH where no defense commit
PR BH Traps	A category of pick & roll BH where the BH gets trapped
PR Roll Man	A category of pick & roll BH where the roll score
Post-Up Defense Commits	A category of Post-Ups where the defense commits
Post-Up Hard Double Team	A category of Post-Ups where a double team comes
Post-Up Single Covered	A category of Post-Ups where no defense commits
Spot Ups	Number of possessions with a Spot-Up play
Total Points	Total Number of Points for the game
Win	1 indicating a win and 0 indicating a loss

1.2 Sets

This dataset contains the number of possessions for every way the Offense sets up.

Table 1.2: Features & Descriptions of Sets Dataset (Source: Synergy Database)

Features	Description
ID	a unique ID to reference the games (row data)
Team	Name of the team

Features	Description
Season	the year of the regular season that the game took place
After Time Outs	Number of possessions that are after a time-out
Half Court Set All	Number of possessions when the offense is set
Half Court Set All No Pts	Number of possessions when the offense is set but no pts
Half Court Set All Pts	Number of possessions when the offense is set & pts scored
Half Court Set Vs Zone	Number of possessions when the offense set vs zone defense
Half Court SetVs.Zone Pts	Number of possessions when the offense is set vs zone&pts
Half-Court SetVs.Zone No	Number of possessions when the offense is set vs zone&nopt
Last 4 Seconds	Number of possessions when it is the last 4 seconds
Out Of Bounds	Number of possessions after inbounding from an outofbounds
Out of Bounds End	Number of possessions from an out of bounds from the end
Out of Bounds Side	Number of possessions from an out of bounds on the side
Total Points	Total Number of Points for the game
Win	1 indicating a win and 0 indicating a loss

1.3 Shots

This dataset contains the types of shots that were taken per game.

Table 1.3: Features & Descriptions of Shots Dataset (Source: Synergy Database)

Features	Description
ID	A unique ID to reference the games (row data)
Team	Name of the team
Season	The year of the regular season that the game took place
2FG Attempts	Number of 2PT Field Goal (FG) Attempts
2FG Made	Number of 2PT FG Made
2FG Missed	Number of 2PT FG Missed
3FG Attempts	Number of 3PT FG Attempts
3FG Made	Number of 3PT FG Made
3FG Missed	Number of 3PT FG Missed

Features	Description
All Free Throws	Number of Free Throws
Live Free Throws	Number of Live Free Throws
FG Attempts	Number of FG Attempts
FG Made	Number of FG Made
FG Missed	Number of FG Missed
Guarded Jump Shots	Number of Guarded Jump Shots
Unguarded Jump Shots	Number of Unguarded Jump Shots
Long Jump Shots	Number of 3 Point Shots
Medium Jump Shots	Number of shots from 17 ft to < 3 point line
Short Jump Shots	Number of shots from < 17 ft
Total Points	Total Number of Points for the game
Win	1 indicating a win and 0 indicating a loss

1.4 Transitions

This dataset contains information about transition plays

Table 1.4: Features & Descriptions of Transitions Dataset (Source: Synergy Database)

Features	Description
ID	a unique ID to reference the games (row data)
Team	Name of the team
Season	the year of the regular season that the game took place
All Push Ball	Number of Possesions where the ball is being pushed
Push Ball - Shot Attempt	A category of Push Ball where the ball is being pushed
Push Ball - Turnover	A category of Push Ball where the ball is being pushed
Push Ball to Half Court	A category of Push ball where the ball is being pushed to
Press Offense	Number of Possessions where the offense is being pressed
Transition Offense	Number of Transition plays
Transition Turnover	Number of Transition plays leading to a turnover
Total Points	Total Number of Points for the game
Win	1 indicating a win and 0 indicating a loss

1.5 Player Statistics

Table 1.5: Features & Descriptions of Player Statistics (Source: OUA website)

Features	Description
Game ID	a unique ID to reference the games (row data)
Date	the date that the game took place
Season	the year of the regular season that the game took place
Team	Name of the team
Player	Name of the player
Home	A binary value; 1 indicating Home team, 0 Away team
GP	Total number of games played
MPG	Minutes played per game
PPG	Points per game
PTS	Total Points scored
MIN	Total minutes played
FGM	Number of Field Goals Made for the team
FGA	Number of Field Goals Attempted for the team
Field Goal%	(FGM/FGA) x 100
3PM	Number of 3 Pointers Made by the team
3PA	Number of 3 Pointers Attempted by the team
3Point%	(3PM/3PA) x 100
FTM	Number of FreeThrows Made by the team
FTA	Number of FreeThrows Attempted by the team
FT%	(FTM/FTA) x 100
Assists	Number of Assists the team made
Rebounds	Number of Rebounds the team made
Steals	Number of Steals in the game
Blocks	Number of Blocks in the game
Turnovers	Number of Turnovers by the team

1.6 General Statistics

Table 1.6: Features & Descriptions of General Statistics (Source: OUA website)

Features	Description
Game ID	a unique ID to reference the games (row data)
Date	the date that the game took place
Season	the year of the regular season that the game took place
Team	Name of the team

Features	Description
Home	A binary value; 1 indicating Home team, 0 Away team
FGM	Number of Field Goals Made for the team
FGA	Number of Field Goals Attempted for the team
Field Goal%	(FGM/FGA) x 100
3PM	Number of 3 Pointers Made by the team
3PA	Number of 3 Pointers Attempted by the team
3Point%	(3PM/3PA) x 100
FTM	Number of FreeThrows Made by the team
FTA	Number of FreeThrows Attempted by the team
FT%	(FTM/FTA) x 100
Assists	Number of Assists the team made
Rebounds	Number of Rebounds the team made
Steals	Number of Steals in the game
Blocks	Number of Blocks in the game
Turnovers	Number of Turnovers by the team
Points off Turnovers	Number of Points made off Turnovers by the team
Points in the Paint	Number of Points made in the Paint by the team
2nd Chance Points	Total 2nd Chance Pts for the team
Bench Points	Number of Pts made by the bench players for the team
Fastbreak Pts	Number of Fastbreak Pts by the team
Largest Lead	The Largest Lead made by the team
Time of Largest Lead	The time of the team's Largest Lead
Win	A binary value; 1 indicating a win, 0 loss
Winner 1st Qtr Pts	The number of points the scored in the 1st qtr
Winner 2nd Qtr Pts	The number of points scored in the 2nd qtr
Winner 3rd Qtr Pts	The number of points scored in the 3rd qtr
Winner 4th Qtr Pts	The number of points scored in the 4th qtr
OT Pts	The number of points scored in overtime

Chapter 2

Exploratory Data Analysis

Exploratory Data Analysis is valuable to data projects because it helps in understanding the data, making sure it is worth investigating, and checking for anomalies. The raw data needs to be validated to ensure that the data set was collected without errors.

2.1 Distribution/Variation of Variables

Distributions are often described in terms of their density or density functions.

Density functions are functions that describe how the proportion of data or likelihood of the proportion of observations change over the range of the distribution. Certain analyses require certain distributions, and if they require all variables to be independently and identically distributed, then standardization will need to be used.

2.1.1 Play Types

Below are basic summary statistics of the Play Types dataset, i.e. the minimum, quartiles, mean, median, and maximum of all the variables. In order to best interpret this data, the reader should refer to Table 1 in section 1.1 where each of the below features and their descriptions are given.

On average, there are 92.05 possessions ("Possessions" highlighted below) per game, among all 1452 regular season games in the dataset. The Spot-Up is the playtype with the highest average (i.e. most frequent during a game) of 22.35 Spot-Ups per game. A Spot-Up is when a player is set in a position to shoot and gets the ball to take the shot. Typically, this is a player waiting at the 3-point line. An Off-Screen possession results from an offensive player getting

the ball when a screen was set by one of their teammates allowing them to be open for a pass. It is important to note these two types of possessions can never happen simultaneously, as a Spot-Up requires no screen being used before the player catches the ball. Examples of a player spotting up are: standing in the corner before catching-and-shooting, relocating to the 3-point line, or fading to the corner and getting the ball on a kick out. These possessions are not just catching and shooting. They can be catching-and-shooting, but attacking a close-out by dribbling into a pull-up, dribbling into a floater, or driving to the rim. It is worthwhile to analyze this playtype as it has the highest frequency among games, and thus coaches improving Spot-Up techniques can be used to a team's advantage.

Summary Statistics for the variables in the Play Types Dataset. Note: an asterisk denotes a factor variable.

mean

sd

median

min

max

range

TotalPoints

77.4400826

13.5884622

78.0

36

125

89

Win*

1.5000000

0.5001723

1.5

1

2

1

Season*

2.6033058

1.1111564

3.0

1

4

3

AllIsolation

8.6053719

4.5569282

8.0

0

27

27

AllOffensiveRebounds

11.0723140

4.0369936

11.0

2

26

24

AllP.RBallHandler

19.0847107

7.3219072

18.0

2

43

41

Possessions

92.0516529

7.6249584

92.0

52

137

85

AllPost.Up

8.4531680

5.2572562

8.0

0

34

34

Cuts

7.1122590

3.4954634

7.0

0

22

22

Handoffs

2.5716253

2.1355756

2.0

0

14

14

Isolation.DefenseCommits

2.6508264

2.0962322

2.0

0

17

17

Isolation.SingleCovered

5.9545455

3.5539555

5.0

0

22

22

MiscellaneousPossessions

6.7520661

3.2296379

6.0

0

20

20

OffScreens

4.0378788

2.7003990

4.0

0

16

16

Off.Reb..PutBacks

5.9035813

2.9562047

6.0

0

19

19

Off.Reb..ResetOffense

5.1687328

2.4915654

5.0

0

15

15

P.RBallHandler.DefenseCommits

10.9931129

5.0872424

11.0

0

32

32

P.RBallHandler.SingleCovered

7.7217631

4.2227431

7.0

0

28

28

P.RBallHandler.Traps

0.3698347

0.8604964

0.0

0

7

7

P.RRollMan

3.1666667

2.3649850

3.0

0

13

13

Post.Up.DefenseCommits

1.6763085

1.7005016

1.0

0

10

10

Post.Up.HardDoubleTeam

1.4407713

1.8879549

1.0

0

15

15

Post.Up.SingleCovered

5.3360882

3.8232836

5.0

0

25

25

SpotUps

22.3519284

5.7683973

22.0

4

44

40

Transitions

18.0172176

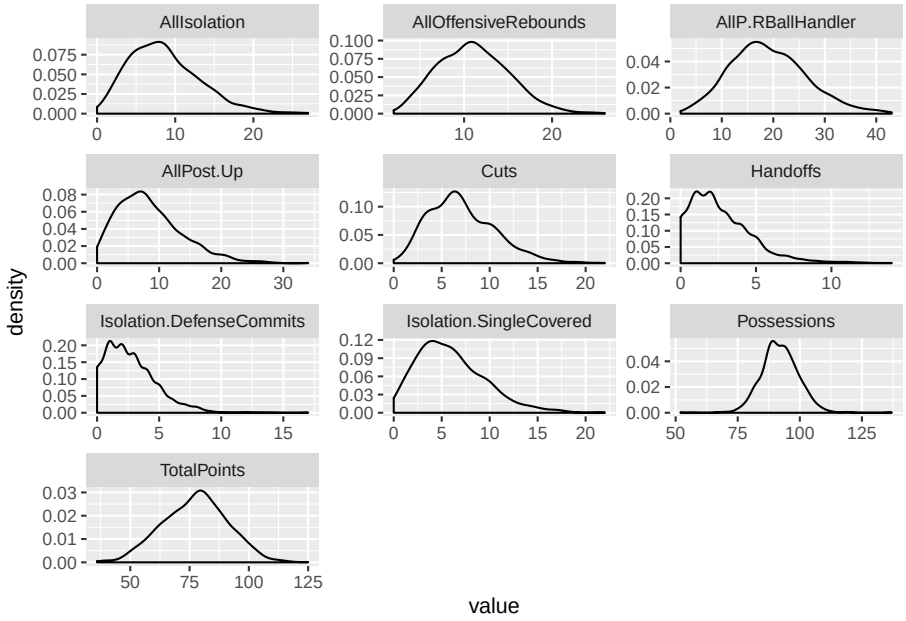
6.1807470

17.0

3

44

41



The distributions of most of the Isolation, Post-Up and Pick and Roll plays are skewed to the right, along with Handoffs, Offscreens and Miscellaneous Posssessions. The rest of the plays are approximately normal.

Note: There is a difference in number of games per season because the number of games played per season increased from 19-20 games to 23-24 games in 2017/2018.

2.1.2 Outliers

An outlier is defined as a sample or event that is very inconsistent with the rest of the data set. However, in sports outliers are not due to measurement errors, they are due to teams playing differently against other teams. Instead, it would be better to average the data and aggregate by team and season.

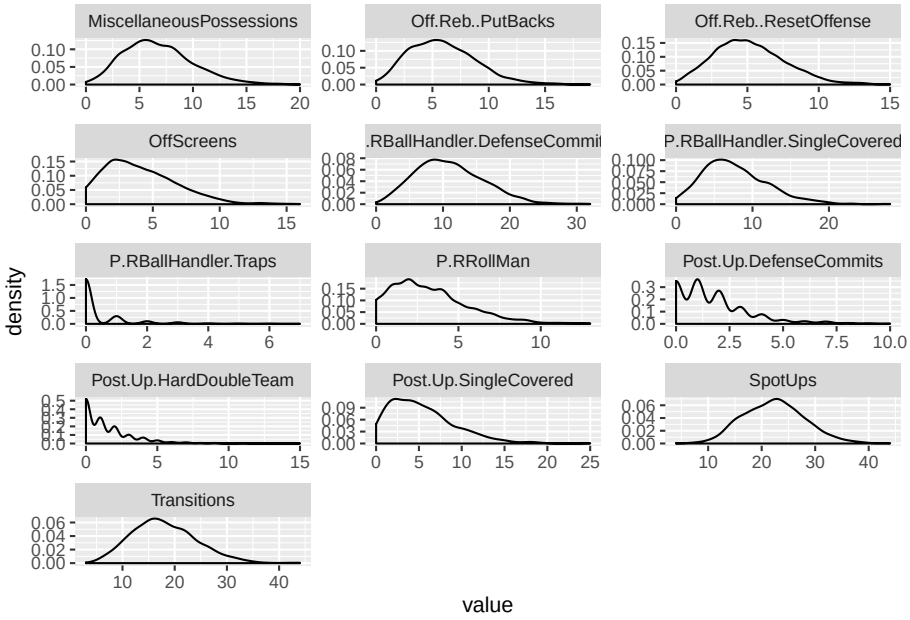
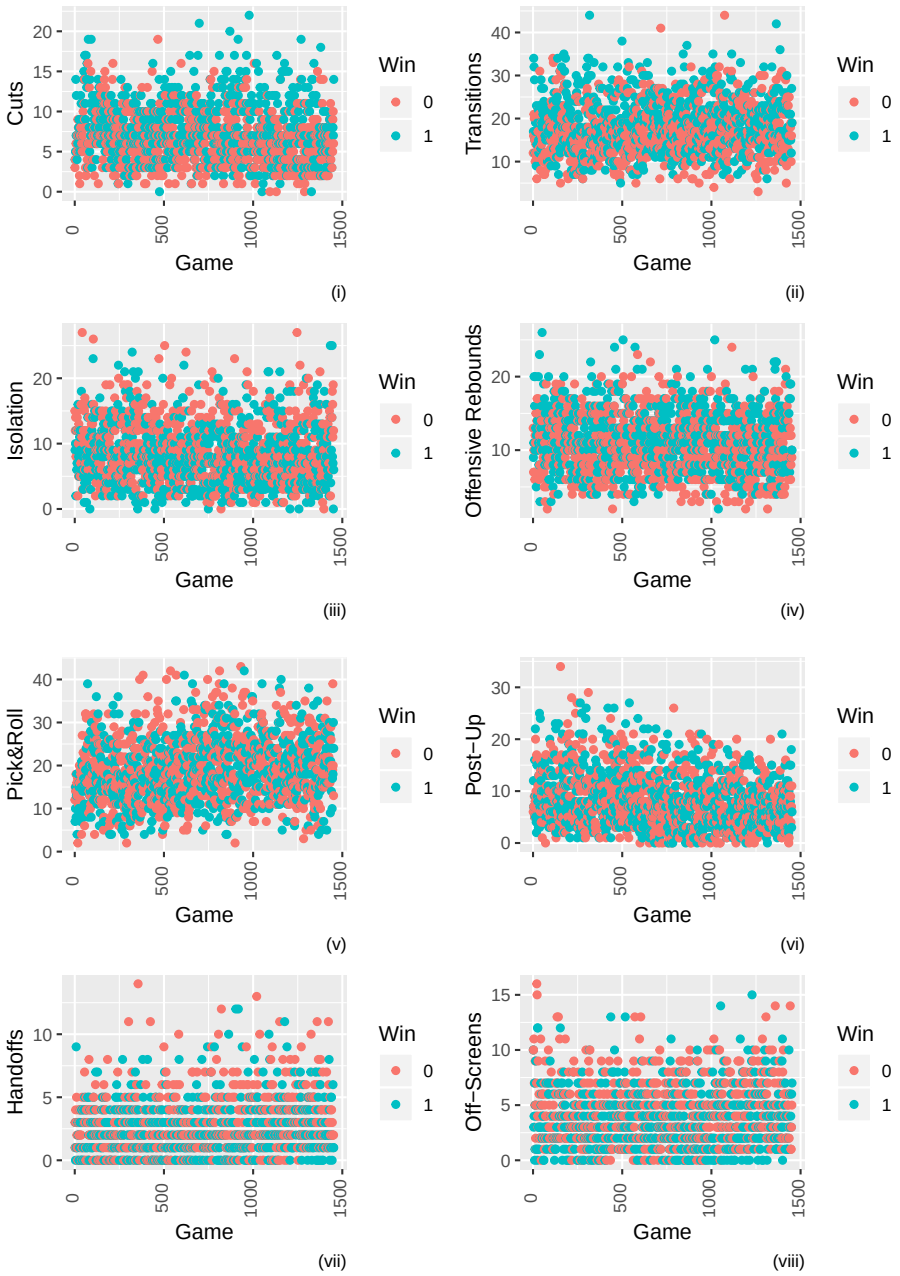


Figure 2.1: Distribution of PlayTypes Features.

2.1.3 Win/Loss Associations

2.1.4 Covariation



There is no clear pattern of any individual play type in respect to wins. This

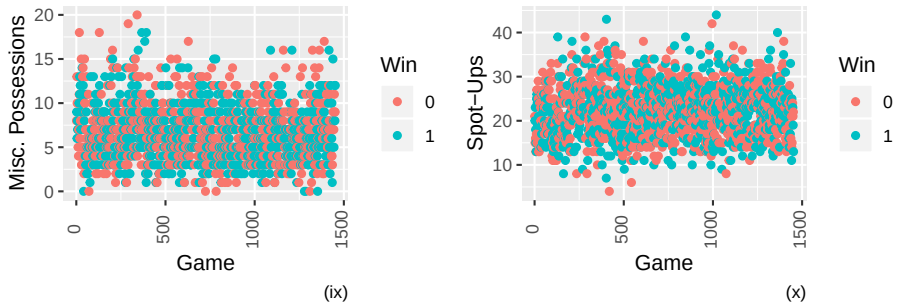


Figure 2.2: Scatterplots of certain Play Types vs. Wins (1) or Losses (0)

makes sense since different teams have different styles of play and have to adjust to their opponents' style of play. It would make more sense to see the differentials for each game. For instance, if a team is not as tall as another team, the taller team may want to post-up more since they would have the advantage. This advantage may make the team more likely to win.

2.1.5 Sets

Below are the basic summary statistics of the Sets dataset which shows the number of times a team sets up their offense and where and when they do. Again, the reader can refer to Table 2 in section 1.2 for the features and their associated descriptions. It may seem like there is an anomaly with the half-court vs zone variables but this is due to zone defense not being a popular defensive style in the league so when a team plays zone defense for the entire game then the opposing team will have to set their offense against it. We can see that zone defenses have right skewed distributions which further shows that zone defense is not a popular defensive style in U Sports Basketball.

Summary Statistics for the variables in the Sets Dataset. Note: an asterisk denotes a factor variable.

mean

sd

median

min

max

range

AfterTimeOuts.ATO.

8.637741

2.0509452

9.0

1

17

16

HalfCourtSetAll

74.034435

7.3461373

74.0

40

113

73

HalfCourtSetAll.NoPts

46.580579

7.0841925

46.0

24

73

49

HalfCourtSetAll.Pts

27.453857

5.2931895

27.0

11

48

37

HalfCourtSetvs.Zone.NoPts

2.807851

5.7575283

1.0

0

46

46

HalfCourtSetvs.Man

69.700413

10.8398110

71.0

6

113

107

HalfCourtSetvs.Man.NoPts

43.772727

8.7634901

44.0

4

71

67

HalfCourtSetvs.Man.Pts

25.927686

5.7791545

26.0

1

45

44

HalfCourtSetvs.Zone

4.334022

8.5731538

1.0

0

77

77

HalfCourtSetvs.Zone.Pts

1.526171

3.0937572

0.0

0

32

32

Last4Sec.ofShotClock

7.323003

3.4618639

7.0

0

20

20

OutofBounds

9.828512

3.1749254

10.0

1

23

22

OutofBounds.End.

5.244490

2.4351274

5.0

0

15

15

OutofBounds.Side.

4.584022

2.2218062

4.0

0

12

12

TotalPoints

77.440083

13.5884622

78.0

36

125

89

Win

0.500000

0.5001723

0.5

0

1

1

Season*

2.603306

1.1111564

3.0

1

4

3

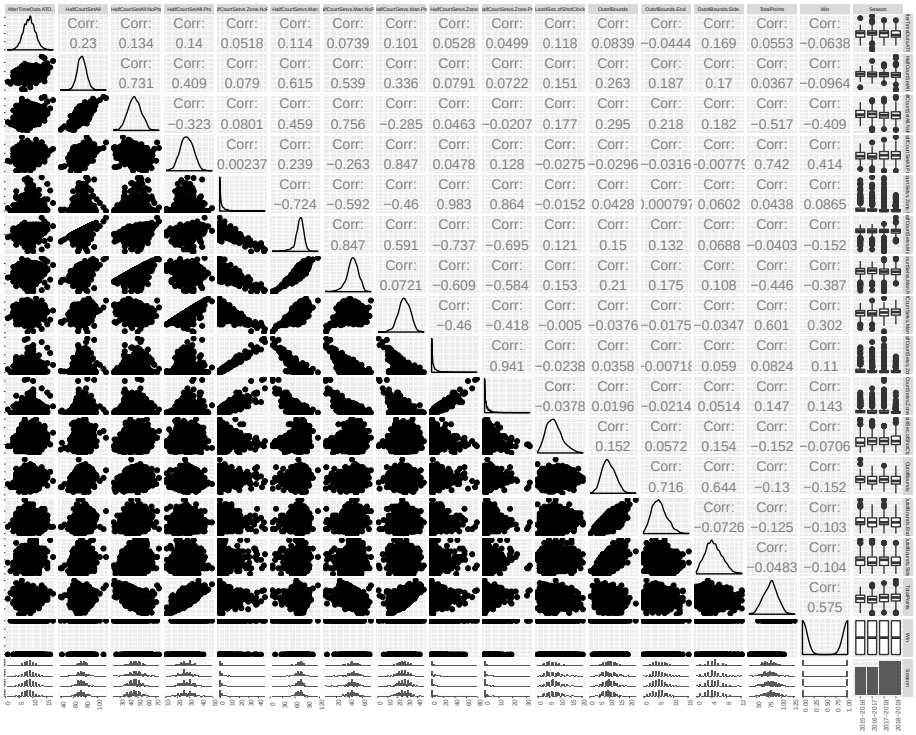


Figure 2.3: Plot Matrix of Sets Dataset.

2.1.6 Shots

Below are summary statistics of the Shots dataset (features and associated description are given in Table 3 in section 1.3). From this we can see that on average, teams take more guarded shots than unguarded shots. Teams also take more long jump shots on average compared to short or medium jump shots. The average FG% from all teams from all 1488 games in the dataset is $27.75/68.1 = 40.75\%$. Teams on average attempt 25 3-Pointers and make about 8 per game which gives an average 3FG% of 32%; 2-Pointers have a higher efficiency on average because they are easier to score. Total Points are negatively correlated to guarded jump shots, short jump shots and medium jump shots, and are positively correlated to long jump shots (3 Pointers). It is self-explanatory that total points are negatively correlated to guarded shots as these have a higher likelihood of being missed. On the other hand, it is interesting to note that teams with players that take more short and medium jump shots as opposed to long shots have less total points, while teams with players taking more long jump shots have more total points. This shows that players with good 3-point shooting efficiency are highly valuable to a team and may in fact be an important factor to a team's season performance.

Summary Statistics for Play Type Variables. Note: an asterisk denotes a factor variable.

mean

sd

median

min

max

range

X2FG.Attempts

43.172865

7.9905014

43.0

19

76

57

X2FG.Made

19.894628

5.2749932

20.0

5

37

32

X2FG.Missed

23.278237

6.1437855

23.0

6

46

40

X3FG.Attempts

25.135675

6.5643332

25.0

8

47

39

X3FG.Made

7.883609

3.2277794

8.0

0

23

23

X3FG.Missed

17.252066

5.0190364

17.0

4

40

36

All.Free.Throws

19.064738

7.0358888

18.0

0

44

44

FG.Attempts

68.308540

7.8644519

68.0

40

102

62

FG.Made

27.778237

5.7206141

28.0

12

51

39

FG.Missed

40.530303

7.0999397

40.0

16

68

52

Guarded.Jump.Shots

12.511708

5.5112691

12.0

1

31

30

Live.Free.Throws

10.068870

3.6851550

10.0

0

23

23

Long.Jump.Shots..3.point.shots.

25.351240

6.5999639

25.0

8

48

40

Medium.Jump.Shots..17..to..3.point.line.

4.294766

2.8733717

4.0

0

19

19

Short.Jump.Shots…17..

4.687328

2.9039348

4.0

0

16

16

Total.Points

77.440083

13.5884622

78.0

36

125

89

Unguarded.Jump.Shots

8.913223

4.8063152

8.0

0

27

27

Win

0.500000

0.5001723

0.5

0

1

1

Season*

2.603306

1.1111564

3.0

1

4

3

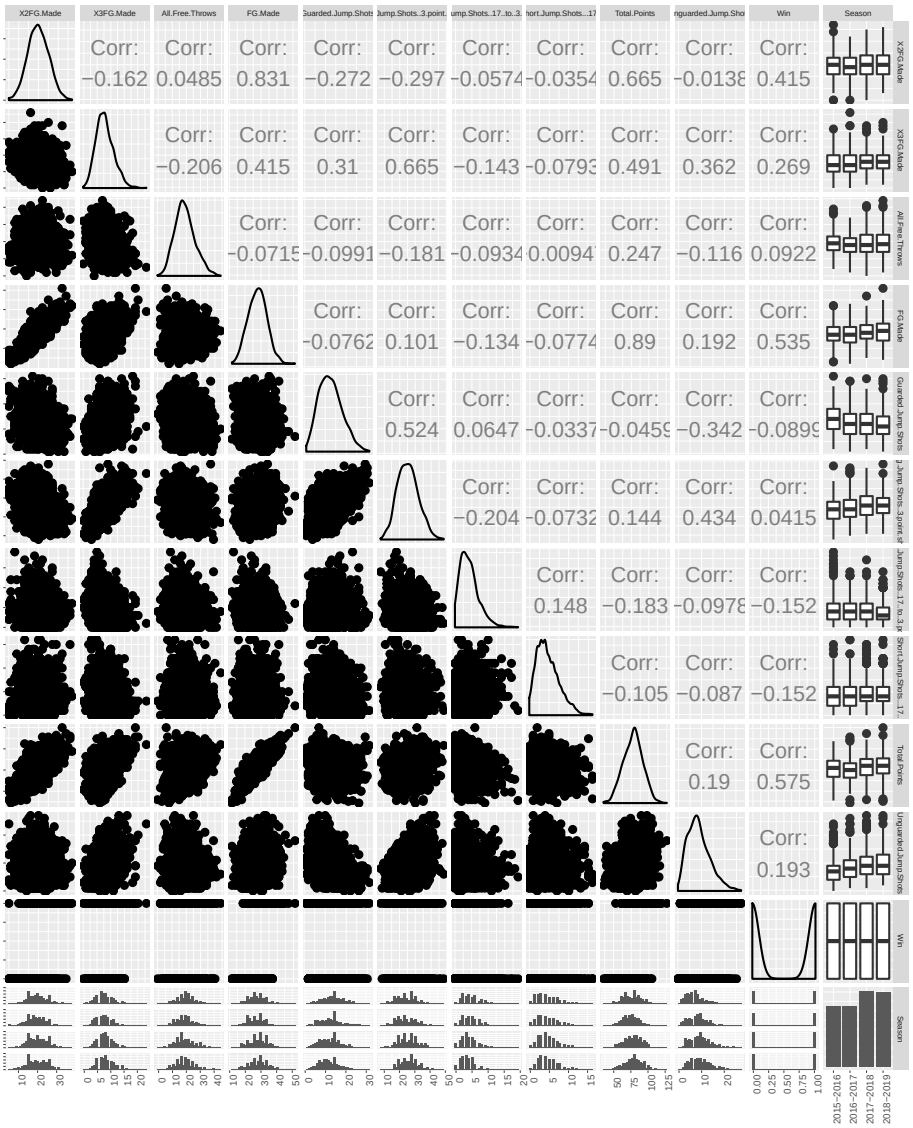


Figure 2.4: Plot Matrix for Shots Dataset

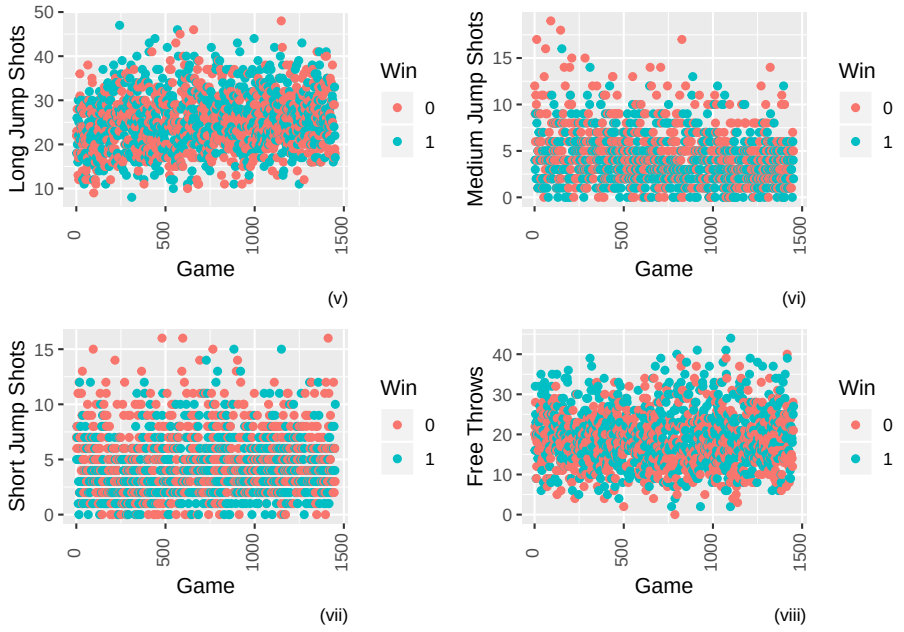
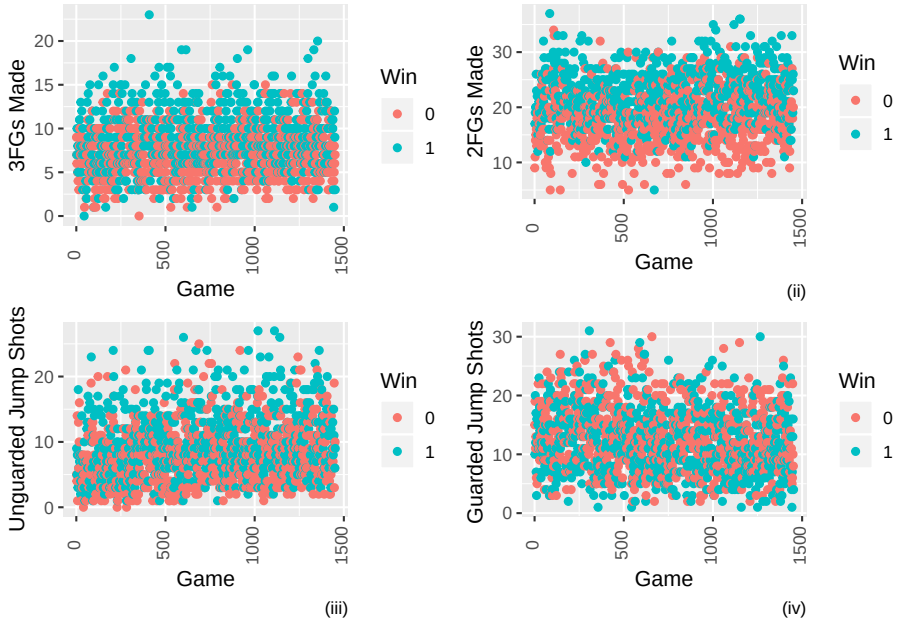


Figure 2.5: Comparing Shot Types vs. Wins(1) or Losses(0)

2.1.7 Visualizations



From the above figure, we can see that more unguarded shots (iii) is more highly associated to wins compared to guarded shots (iv). In this figure we can see that taking a lower number of medium jump shots (vi) contribute to more wins as opposed to the other types of shots (v & vii) that are taken.

2.1.8 Transitions

Below are summary statistics of the Transitions dataset (features and associated descriptions are given in Table 4 in section 1.4). Total Points is most positively correlated to Transition Offense with 0.36 where Transition Offense occurs when a team gains possession of the ball and quickly pushes it to the opposing team's basket. Total Points is most negatively correlated to Press Offense. Press Offense is when the offense (the team having possession of the ball) is being pressed by the other team, i.e. they are being defensively pressured in which members of the defense cover their opponents throughout the court and not just near their own basket. Being pressured would make it harder to score, thus why it is the most negatively correlated to points. The outliers (shown in the boxplots) are all on the upper tails and may be due to the pace of game having a big variance. For example, a team may have a higher Transition Offense rate when the pace of the game is fast, but if the pace is slow, they may not transition from defense to offense as often. The outliers should not be removed from the dataset since they are not measurement errors and provide useful information where the data points largely deviate from the average.

2.1.9 General Statistics

2.1.10 Home Vs. Away

The distributions for the home variables vs the away ones are very similar, however there is a slight difference between the Field Goal Percentage.

Table 2.1: Home Shooting Efficiency vs. Away Shooting Efficiency

	Average	Statistic
Away	0.4092	FG%
Home	0.4228	FG%
Away	0.3123	3FG%
Home	0.3281	3FG%

There is a very slight difference between the home and away field goal percentages but does this mean that there is a home court advantage?

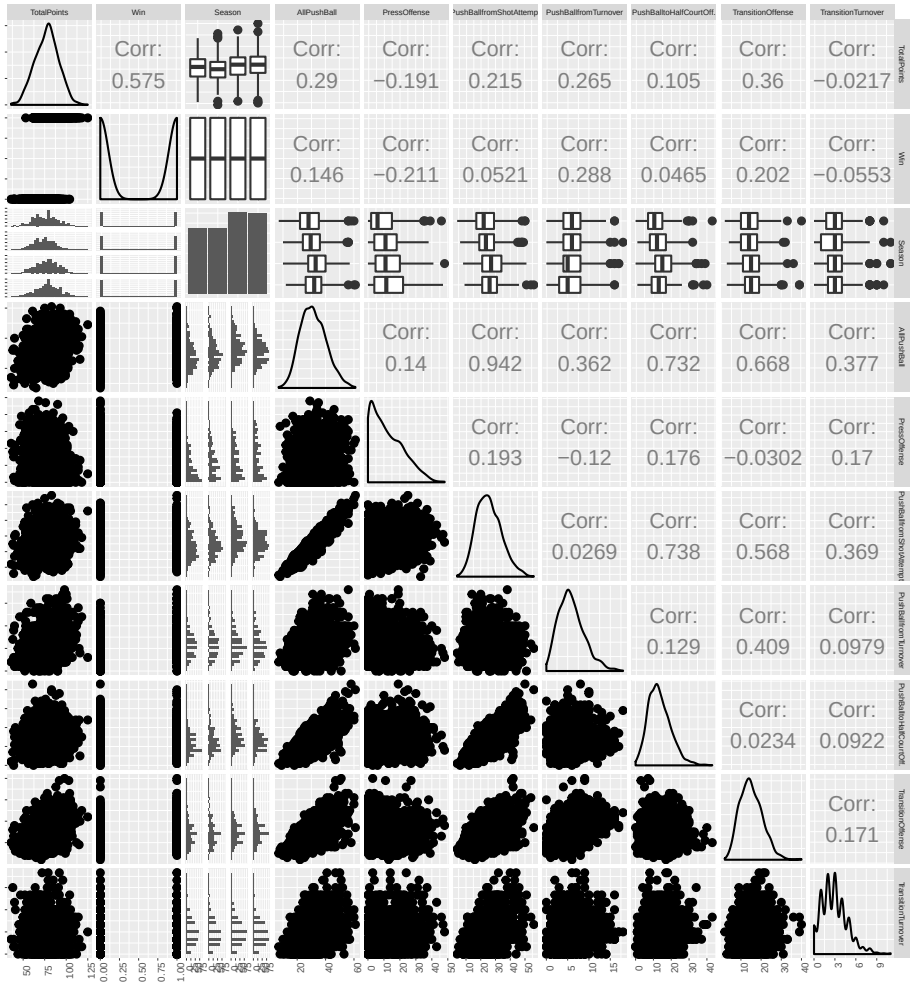


Figure 2.6: Plot Matrix of the Transitions Dataset

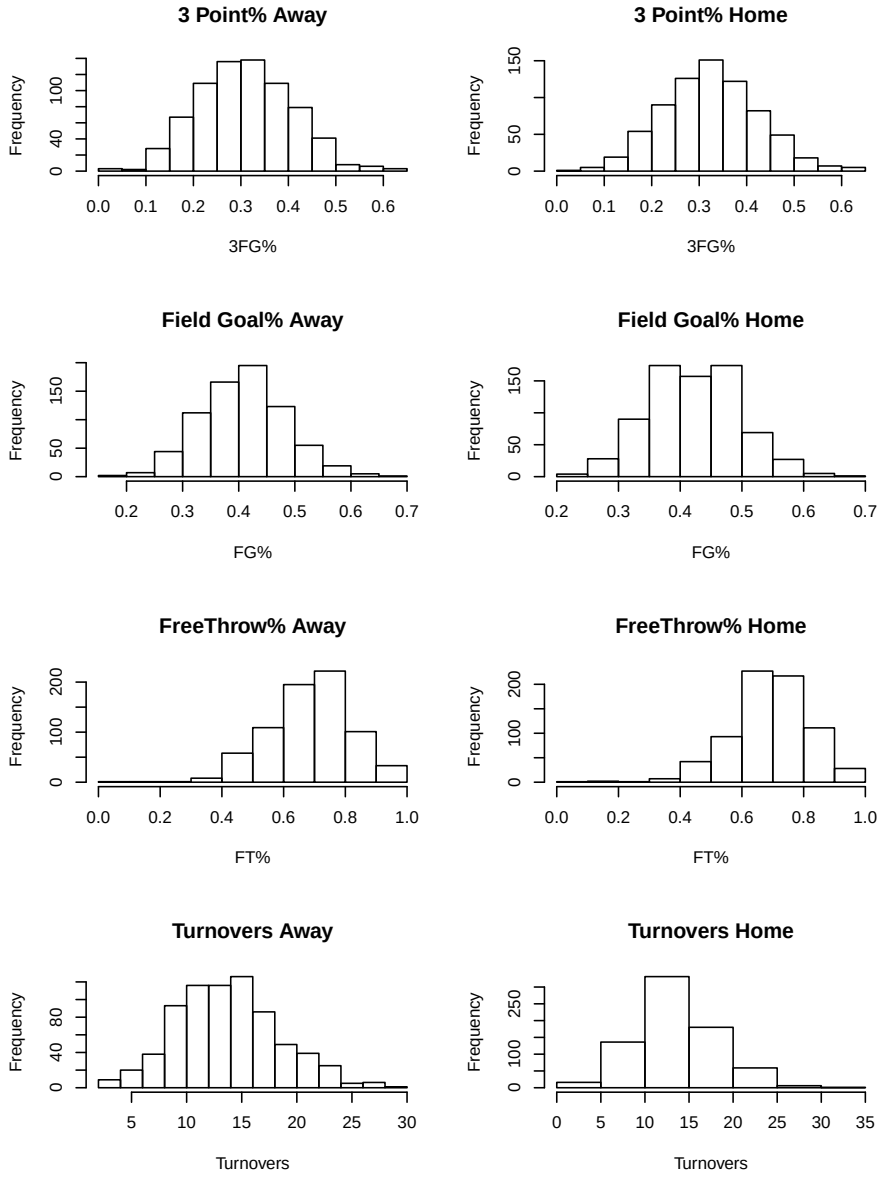


Figure 2.7: Distribution of Variables; Away vs. Home

Table 2.2: Home Wins vs. Away Wins

Home Wins	Away Wins
402	328

This shows there is a difference between the number of times a home team wins compared to an away team.

2.1.11 Risk Ratios and Odd Ratios

2.1.11.1 What are Risk Ratios and Odd Ratios

Risk Ratio (RR) or Relative Risk is a measurement often used in epidemiology. It is used to estimate the outcome between factors and outcomes. In our case we will use this measurement to see whether there is a statistically significant difference between teams playing at home versus away. A risk ratio of 1 means there is no difference, greater than 1 means there is a higher chance of winning if the team is playing at home, and less than 1 means the opposite [4]. An Odds Ratio (OR) is a ratio of ratios. It also quantifies the strength of the association between two events. If the odds ratio equals 1 then the odds of the events are the same. If the odds ratio is greater than 1 then the events are correlated in the sense that if compared to the absence of the second event, the presence of the second raises the odds of the first event, and symmetrically the presence of the first event raises the odds of the second event. In our case we will obtain both measurements to see the strength of association between teams playing at home versus teams playing away.

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Win
Comparing : Home vs. Away

     Win Lose    P(Win) 95% conf. interval
Home 402  328    0.5507    0.5144   0.5864
Away 328  402    0.4493    0.4136   0.4856

                                  95% conf. interval
            Relative Risk: 1.2256    1.1049   1.3595
         Sample Odds Ratio: 1.5021    1.2222   1.8462
Conditional MLE Odds Ratio: 1.5017    1.2156   1.8562
    Probability difference: 0.1014    0.0501   0.1519

             Exact P-value: 0.0001
```

Team	Season	FT%	FG%	3P%	Turnovers Per Game	Assists Per Game	Rebounds Per Game	Points Per Game
Algoma	2018-19	3.56%	2.75%	0.92%	-3.17	0.00	-6.25	3.75
Brock	2018-19	-1.91%	1.93%	0.82%	2.50	3.17	9.08	6.00
Carleton	2018-19	8.91%	3.47%	3.30%	-0.38	6.19	-0.61	8.16
Guelph	2018-19	1.61%	4.36%	-0.36%	1.25	2.17	0.83	5.08
Lakehead	2018-19	5.58%	-2.48%	-0.85%	0.58	3.25	0.25	1.00
Laurentian	2018-19	4.60%	4.75%	8.18%	-2.65	4.60	6.11	9.74
Laurier	2018-19	3.62%	-5.61%	-4.16%	-0.42	1.00	2.00	-6.00
McMaster	2018-19	2.07%	-0.63%	3.56%	5.00	4.33	4.33	0.75
Nipissing	2018-19	3.97%	1.19%	5.69%	0.80	3.17	-7.29	0.63
Ottawa	2018-19	-1.64%	6.44%	11.20%	-2.36	5.08	3.92	12.11
Queen's	2018-19	-2.80%	1.53%	-2.74%	-1.10	5.51	0.09	5.02
Ryerson	2018-19	-3.14%	-0.05%	-2.86%	2.62	1.32	10.64	2.76
Toronto	2018-19	1.87%	5.59%	-1.63%	-2.60	0.92	3.36	10.89
Waterloo	2018-19	-5.19%	-2.70%	3.10%	0.32	0.60	-0.57	-3.71
Western	2018-19	12.32%	2.91%	1.78%	-2.75	-2.17	-3.67	1.08
Windsor	2018-19	0.13%	4.88%	3.09%	-1.11	6.39	0.50	8.01
York	2018-19	0.72%	2.44%	4.37%	-2.41	-2.90	-2.94	3.64

Figure 2.8: Difference of Home Statistics vs. Away Statistics of the 2018-2019 Season for each team.

```
        Asymptotic P-value: 0.0001
------------------------------------------------------
```

The probability of winning at home is 55% whereas the probability of winning away is 45%. The Sample Odds Ratio tells us that odds of a team winning is 1.5 higher given they are playing at home compared to playing away. The Relative Risk tells us that home teams have 1.22 times the 'risk' of winning compared to away teams.

A coach may be more interested in which teams in particular play better at home, and how much better they play.

2.1.11.2 Home vs. Away by Team

Above is a table of the every team from the 2018-2019 season where the Home statistics are all subtracted by the Away Statistics, i.e. the statistics of a team when they were playing at home subtracted by statistics when they were playing away. A positive number indicates that the team performed better at home (except for turnovers). For example, Carleton shot their free throws 8.91% higher at home.

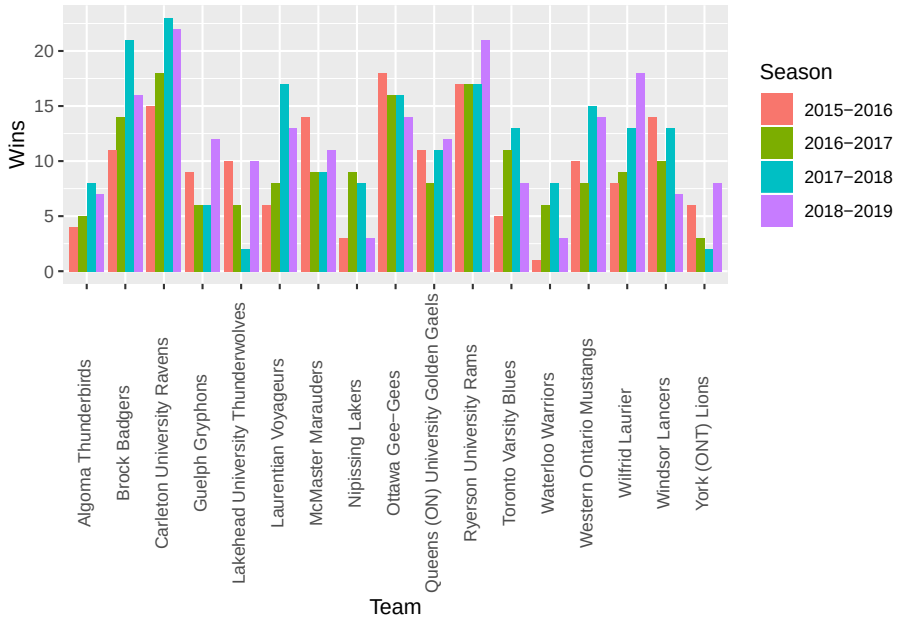
2.1.11.2.1 Insights

The top 3 teams that shot their free thows better at home are Western (12.32%), Carleton (8.91%), and Lakehead (5.58%). The top 3 teams that shot field goals better at home are Ottawa (6.44%), Toronto (5.59%), and Windsor (4.88%). The top 3 teams that shot 3 pointers better at home are Ottawa (11.20%), Laurentian (8.18%), Nipissing (5.69%). The top 3 teams that turnover the ball the least when playing at home are Algoma (-3.17), Western (-2.75), and Laurentian (-2.65). The top 3 teams that rebound the ball more at home are Ryerson (10.64), Brock (9.08), and Laurentian (6.11). The top 3 teams that scored more points at home are Ottawa (12.11), Toronto (10.89), and Laurentian (9.74). On average, the teams turned over the ball 6 less times at home,

2.1.11.2.2 Conclusion

In conclusion, many teams benefit from playing at home, and different teams excel differently. According to a Bleacher Report study [5], referee bias and the psychological impact of playing at home are two of the biggest factors of why there is a large difference between home and away statistics. Studies have show that when a crowd is vocal, it impacts the way referees call a game. Also, referees have historically favored home teams. In addition, the psychological impact of playing at home is a self-sustaining placebo effect: Home-court advantage gives the home team an edge simply because players believe that it does.

2.1.12 Wins Per Season



The above shows that Brock, Carleton, Laurentian, UofT, and Western all steadily improved and peaked at the 2017-2018 season. The Ryerson Rams stayed consistent and peaked 2018-2019 season. There are few teams that are consistently not winning more than 10 games a season such as Algoma, Nipissing and York.

2.2 Correlations

The table below gives the correlations between different Play Types and Total Points scored in a game. Note that a negative number represents a negative correlation between the two features while a positive number represents a positive correlation. A correlation measurement closer to 0 represents a non-linear relationship as opposed to a correlation measurement further from 0.

Table 2.3: Correlation between Play Types and Total Points scored in a game.

Play Type	Correlation to Total Points
All Isolation	-0.046532124
All Offensive Rebounds	**0.154450763**
All PR Ball Handler	0.042544034
All Post-Up	-0.040890549
Cuts	**0.227413916**
Handoffs	-0.017105738
Isolation Defense Commits	-0.045338194
Isolation Single Covered	-0.032922237
Miscellaneous Possessions	-0.075670507
OffScreens	**-0.135119217**
Offensive Rebound Putback	0.154161317
Offensive Rebound Reset Offense	0.067340925
PR Ball Handler Defense Commits	0.053122283
PR Ball Handler Single Covered	0.013881865
PR Ball Handler Traps	-0.020176743
PR Roll Man	0.101940646
Post Up Defense Commits	-0.056702755
Post Up Hard Double Team	-0.090199983
Post Up Single Covered	0.013534056
Spot Ups	-0.007428537
Transitions	**0.317687812**

The plays that are most positively correlated to total points are transitions, cuts, and offensive rebounds. This could mean that transitions, cuts and offensive rebounds contribute to the most points compared to all other plays. The play that is most negatively correlated to total points is offscreens.

To account for outliers and since some teams have played more a game or two more than others, the dataset was transformed by averaging the statistics per game per season, and the wins were summed.

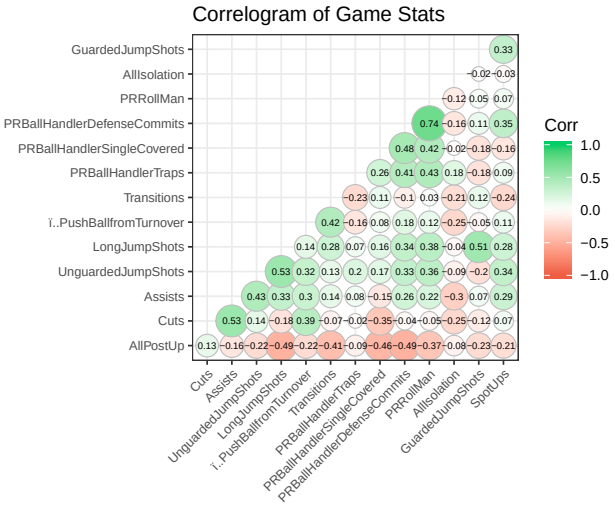
Table 2.4: Correlation between Game Statistics and Number of Wins.

Features	Correlation to Number of Wins
Press Offense	-0.175526612
Push Ball From Shot Attempt	0.091290517
Push Ball From Turnover	**0.484877383**
Push Ball to Half Court	0.115357730
Free Throws	-0.092037012
Guarded Jump Shots	-0.153775302
Unguarded Jump Shots	**0.592888473**
Long Jump Shots	**0.331422037**
Medium Jump Shots	-0.365731181
Short Jump Shots	-0.197920494
Cuts	**0.373973357**
Handoffs	-0.072249207
Isolation Single Covered	-0.354371026
Isolation Defense Commits	-0.006820286
Miscellaneous Possessions	-0.406512329
OffScreens	-0.280720057
Offensive Rebound PutBack	0.166641468
Offensive Rebound Reset	**0.368757978**
PR Ball Handler Defense Commit	**0.338107184**
PR Ball Handler Single Covered	0.072595284
PR Ball Handler Traps	0.115888712
PR Roll Man	**0.311420152**
Post Up Defense Commits	-0.171605486
Post Up Hard Double Team	-0.067993980
Post Up Single Covered	-0.183923948
Spot Ups	**0.247910273**
Transitions	**0.231836305**
Assists	**0.653151380**
Blocks	**0.337827582**
Steals	**0.490803284**
Total Rebounds	**0.447741075**
Turnovers	-0.319146688

2.2.1 The most positively correlated variables to wins

The most positively correlated variable to number of wins is assists with a correlation of 0.65. Next to that are unguarded jump shots with a correlation of 0.59. The play types that are positively correlated to wins are Offensive Rebound Reset Offense, P&R Ball Handler Defense Commits, P&R Roll Man, Cuts, Transitions and Spot-Ups. Offensive Rebound Reset Offense gives the

team another chance to score, P&R Ball Handler Defense Commits would leave a man open to score, P&R Roll Man can lead a man to an unguarded shot and same for Cuts, Transitions and Spot-Ups. The shot types with the highest correlation are the Long Jump Shot (3 Pointers), and of course, the Unguarded Jump Shots. Furthermore, Push Ball from Turnover is also highly correlated with wins which makes sense because if another team turns over the ball then they wasted a possession and the other team is able to score (most usually in a fastbreak). The general statistics that are most positively correlated to wins are assists, blocks, steals and rebounds. Blocks, steals and rebounds create more possessions to teams while creating less for the other team, i.e. the more you steal, block or rebound the ball, the more chances you have to score while putting your opponent at a disadvantage.



This figure shows that the most positively correlated statistics to unguarded jump shots are long jump shots, P&R Roll Man, Spot-Ups, P&R Ball Handler Defense Commits, and Push Ball from Turnover.

2.2.2 The most negatively correlated variables to wins

The most negatively correlated variable to number of wins is Miscellaneous Possessions with a correlation of -0.40. Miscellaneous Possessions are undefined plays, possibly due to confusion, sloppy play, or bad decisions. The shot types that are negatively correlated are medium jump shots, short jump shots and guarded jump shots. The negatively correlated play types are Isolation Single Covered, Post-Up Single Covered, and OffScreens. This may suggest that these plays are easier to defend or harder to score from. And of course, the most negatively correlated general statistic is turnovers.

Chapter 3

Player Analysis

3.1 Dataset

The data aggregated is scraped from the OUA website. Every box score per game per season is collected and aggregated so that there are player statistics for every season from 2014-15 to 2018-19 (this is because the Player Statistics came from the OUA website which has 2014-15 data).

3.2 The Goal

The goal is to use player statistics to gather insights on how players contribute to the game and how to categorize players using unsupervised learning.

3.3 Data Preparation

The data is a subset of the player data with certain filters on the number of games played and the minutes per game. There are dataframes for every season from 2014-15 to 2018-19 with players that have played at least 15 games of at least 20 minutes per game. The games are regular season games from the U Sports division, Ontario University Athletics conference. All the variables are totals for the season except for PPG (Points per game) and MPG (Minutes per game)

$$W(P^q) = \frac{1}{q}\sum_{k=1}^{q}\frac{1}{n_k}\sum_{i=1}^{n_k} d(x_i, c_k)$$

Figure 3.1: Intra-cluster Inertia formula.

3.4 K-Means Clustering

First the data will be normalized in order to prepare the data for k-means clustering. This is helpful because some statistics have very different ranges e.g. the number of points compared to the number of steals. Therefore the variables will be comparable.

K-Means Clustering is a popular unsupervised machine learning algorithm. The goal of K-Means is to group similar data points in a dataset of unlabeled data. It does this by dividing the data into k clusters where each observation belongs to the cluster closest to the mean (cluster centroid) by using a distance metric (most usually Euclidean distance).

Since K-Means clustering is an unsupervised algorithm, this means that the number of clusters is not known. However, there are techniques that can be used to find an optimal number of clusters such as the gap method, silhouette method, within-cluster sum of squares method, D - index, etc. Different techniques and configurations of the techniques will be used for each season's clustering for finding the optimal number of clusters.

3.4.1 K-Means Results

The technique that will be used to find the optimal number of clusters for the 2014-2015 season is the D-index method (Lebart et al. 2000). The D-index is based on clustering gain on intra-cluster inertia [8]. Intra-cluster inertia can be defined as:

The clustering gain should be minimized. The optimal cluster configuration can be identified by the sharp knee that corresponds to a significant decrease of the first differences of clustering gain versus the number of clusters. This knee or great jump of gain values can be identified by a significant peak in second differences of clustering gain.

In the plot of D-index, we seek a significant knee (the significant peak in D-index second differences plot) around 8, that corresponds to a significant increase of the value of measure. The number of clusters that the method suggests is 8 clusters.

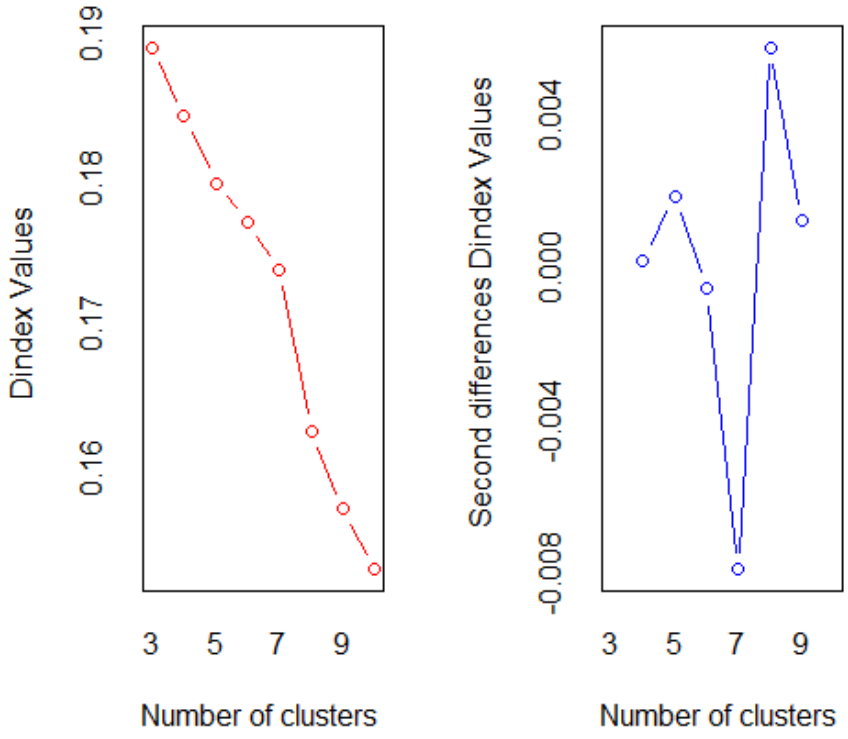


Figure 3.2: Finding the Optimal Cluster using D-index for the 2014-15 season.

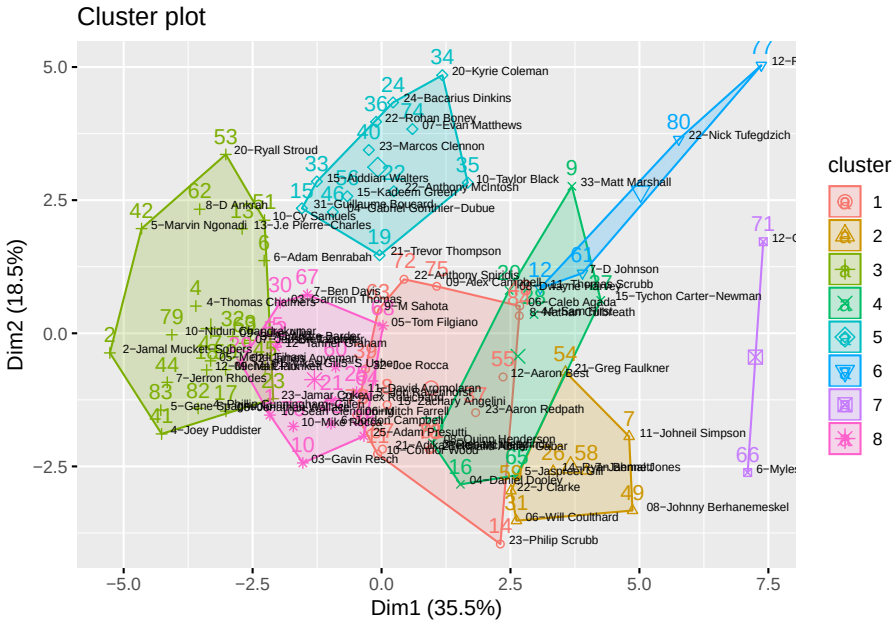


Figure 3.3: Cluster Plot for 2014-15 Season. The axes are the Principal Components where Dim1 is the first PC and Dim2 is the second PC. The first PC explains 35.5% of the data and the second PC explains 18.5%.

Average of each variable for each cluster for the 2014-15 season

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

Cluster 6

Cluster 7

Cluster 8

3PointMade

31.47

40.14

12.77

19.57

2.25

17.25

23.50

19.47

3PointAttempted

81.40

112.57

37.86

64.29

8.92

46.75

74.00

56.47

Assists

38.07

56.14

22.50

47.71

22.17

36.50

59.00

35.20

Blocks

4.33

6.29

3.64

5.86

15.25

21.25

4.50

4.13

DefensiveRebounds

59.60

64.86

44.18

81.86

79.83

128.75

94.00

58.00

FieldGoalMade

84.87

118.43

42.91

87.29

78.50

132.50

149.00

50.53

FieldGoalAttempted

194.00

278.14

104.55

209.29

157.50

264.75

304.50

135.27

FreeThrowsMade

44.80

58.57

20.86

55.71

40.50

64.25

99.50

29.40

FreeThrowsAttempted

59.87

73.57

32.59

75.00

62.83

94.75

137.50

40.67

Minutes

502.93

579.86

387.45

619.14

450.75

547.25

694.00

500.93

OffensiveRebounds

25.67

16.86

19.00

25.86

44.50

52.50

26.50

16.67

PersonalFouls

39.40

35.43

39.50

50.71

46.17

33.50

44.50

44.93

Points

246.00

335.57

119.45

249.86

199.75

346.50

421.00

149.93

Rebounds

85.27

81.71

63.18

107.71

124.33

181.25

120.50

74.67

Steals

19.60

21.86

13.00

23.57

12.50

16.75

27.50

20.53

Turnovers

34.47

40.14

28.55

48.43

30.50

47.25

57.00

34.27

Home

9.53

9.00

9.27

9.14

9.50

9.50

10.50

9.47

GamesPlayed

18.73

18.29

18.00

19.00

18.92

18.75

20.00

18.93

PointsPerGame

13.15

18.39

6.64

13.19

10.59

18.47

21.05

7.95

MinutesPerGame

26.85

31.74

21.53

32.60

23.81

29.20

34.70

26.53

3P%

0.37

0.36

0.30

0.29

0.18

0.37

0.32

0.38

FG%

0.44

0.43

0.42

0.42

0.50

0.50

0.49

0.38

FT%

0.74

0.79

0.66

0.76

0.64

0.70

0.72

0.72

TrueShooting%

0.56

0.54

0.50

0.52

0.54

0.56

0.57

0.49

Players from 2014-15 season with assigned clusters.

Player

Team

Cluster

15-Zachary Angelini

Brock

1

10-Connor Wood

Carleton

1

23-Philip Scrubb

Carleton

1

23-Aaron Redpath

McMaster

1

32-Joe Rocca

McMaster

1

06-Caleb Agada

Ottawa

1

12-Aaron Best

Ryerson

1

21-Adika Peter-McNeilly

Ryerson

1

9-M Sahota

Toronto

1

22-Anthony Spiridis

Western

1

06-Mitch Farrell

Windsor

1

09-Alex Campbell

Windsor

1

21-Khalid Abdel-Gabar

Windsor

1

3-Richard Iheadindu

York

1

8-Nathan Culbreath

York

1

11-Johneil Simpson

Brock

2

14-Ryan Bennett

Laurentian

2

06-Will Coulthard

Laurier

2

08-Johnny Berhanemeskel

Ottawa

2

21-Greg Faulkner

Queen's

2

7-Jahmal Jones

Ryerson

2

22-J Clarke

Toronto

2

2-Jamal Mucket-Sobers

Algoma

3

3-AJ Andre Barder

Algoma

3

4-Thomas Chalmers

Algoma

3

6-Adam Benrabah

Algoma

3

13-J.e Pierre-Charles

Carleton

3

05-Jonathan Wallace

Guelph

3

12-Michel Clark

Guelph

3

23-Jamar Coke

Lakehead

3

09-Luke Allin

Laurier

3

4-Joey Puddister

Nipissing

3

5-Marvin Ngonadi

Nipissing

3

7-Jerron Rhodes

Nipissing

3

01-Vikas Gill

Ottawa

3

05-Mehdi Tihani

Ottawa

3

09-Matt Plunkett

Ottawa

3

10-Cy Samuels

Queen's

3

20-Ryall Stroud

Queen's

3

8-D Ankrah

Toronto

3

07-Jedson Tavernier

Western

3

10-Nidun Chandrakumar

York

3

4-Phillip Cunningham-Gillen

York

3

5-Gene Spagnuolo

York

3

33-Matt Marshall

Brock

4

04-Daniel Dooley

Guelph

4

08-Dwayne Harvey

Lakehead

4

15-Tychon Carter-Newman

Laurentian

4

44-Sam Hirst

Laurentian

4

5-Jaspreet Gill

Waterloo

4

08-Quinn Henderson

Western

4

31-Guillaume Boucard

Carleton

5

21-Trevor Thompson

Guelph

5

22-Anthony McIntosh

Lakehead

5

24-Bacarius Dinkins

Lakehead

5

15-Aiddian Walters

Laurier

5

20-Kyrie Coleman

Laurier

5

10-Taylor Black

McMaster

5

22-Rohan Boney

McMaster

5

23-Marcos Clennon

Nipissing

5

04-Gabriel Gonthier-Dubue

Ottawa

5

15-Kadeem Green

Ryerson

5

07-Evan Matthews

Windsor

5

11-Thomas Scrubb

Carleton

6

7-D Johnson

Toronto

6

12-Rotimi Osuntola

Windsor

6

22-Nick Tufegdzich

York

6

6-Myles Charvis

Waterloo

7

12-Greg Morrow

Western

7

10-Sean Clendinning

Algoma

8

5-Brett Zufelt

Algoma

8

03-Gavin Resch

Carleton

8

21-Alex Robichaud

Lakehead

8

11-David Aromolaran

Laurentian

8

02-James Agyeman

Laurier

8

03-Garrison Thomas

Laurier

8

25-Adam Presutti

McMaster

8

6-Jordon Campbell

Nipissing

8

12-Tanner Graham

Queen's

8

5-S Usher

Toronto

8

3-Jon Ravenhorst

Waterloo

8

7-Ben Davis

Waterloo

8

05-Tom Filgiano

Western

8

10-Mike Rocca

Windsor

8

Each cluster can be categorized as a type of player.

Cluster 1: Efficient Playmakers & Scorers This cluster of players have the most assists and the second most points per game. They have a big defensive impact through the number of steals they get and can control the tempo well and score.

Cluster 2: All-Around Players These players can get rebounds, pass and score well.

Cluster 3: Dominant Big Men These players are the most dominant big men in the league with the most rebounds (defensive and offensive), blocks, and points.

Cluster 4: Smart Catch & Shoot Players These players make the best decisions and turnover the ball the fewest. They do not dribble the ball much and are the most efficient shooters.

Cluster 5: Aggressive Defenders These players are aggressive and foul the most out of all the other clusters. They have a bigger impact on defense since they do not shoot well.

Cluster 6: Role Players These players contribute to many plays and work both offensively and defensively.

Cluster 7: Second Tier Playmakers These players are less dominant playmakers that can still score efficiently.

Cluster 8: Second Tier Small Players These players play small but do not shoot as efficiently as the other players or create as many plays.

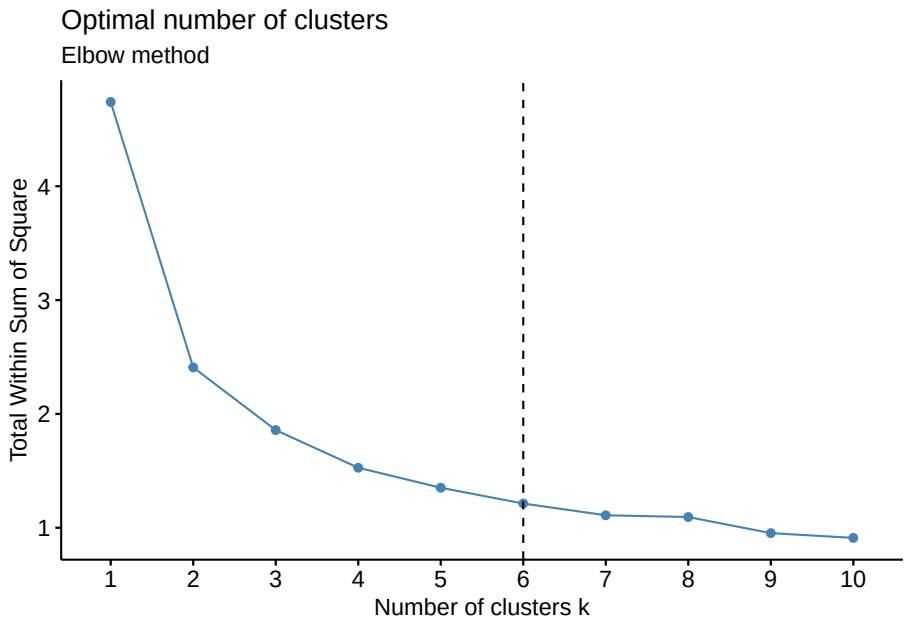


Figure 3.4: Elbow Method for finding the optimal number of clusters for the 2015-16 season

3.4.2 2015-16 Season

For this season, the elbow method [9] will be used to find the optimal number of clusters. The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster does not give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified.

The number of clusters that will be used for this season is 6 for this season. Below are tables to show the average statistics for each cluster and also which players belong to which cluster.

Average of each variable for each cluster for the 2015-16 season

Cluster 1

Cluster 2

Cluster 3

Cluster 4

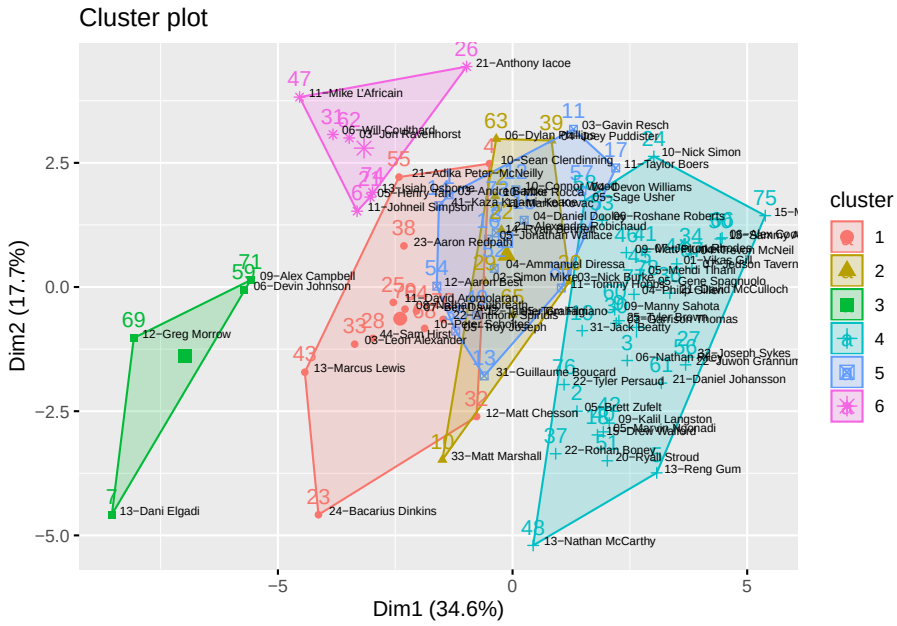


Figure 3.5: Cluster Plot of 2015-16 Season

Cluster 5

Cluster 6

3PointMade

14.85

22.62

18.75

13.16

31.75

44.57

3PointAttempted

48.15

67.88

55.75

41.84

86.00

133.29

Assists

39.38

57.12

43.25

24.91

40.62

52.71

Blocks

7.85

4.75

13.50

7.41

6.44

3.29

DefensiveRebounds

72.00

72.50

111.25

51.62

63.56

68.71

FieldGoalMade

101.38

56.25

150.50

48.56

85.62

111.14

FieldGoalAttempted

236.23

150.62

317.75

119.72

196.88

290.00

FreeThrowsMade

59.38

28.75

99.00

23.69

32.88

55.14

FreeThrowsAttempted

83.46

40.12

130.25

34.47

43.44

74.00

Minutes

564.00

620.00

631.50

419.31

490.81

639.14

OffensiveRebounds

30.77

25.75

44.50

21.50

18.19

14.57

PersonalFouls

44.92

48.88

47.25

39.56

38.44

35.57

Points

277.00

163.88

418.75

133.97

235.88

322.00

Rebounds

102.77

98.25

155.75

73.12

81.75

83.29

Steals

23.85

25.00

30.25

15.72

18.88

24.14

Turnovers

49.46

40.50

58.00

26.00

34.50

49.57

Home

9.69

9.62

9.75

9.31

9.25

9.71

GamesPlayed

19.08

19.38

19.00

18.38

18.25

19.43

PointsPerGame

14.58

8.47

22.02

7.34

12.99

16.58

MinutesPerGame

29.60

31.98

33.18

22.87

26.96

32.90

3P%

0.26

0.32

0.31

0.28

0.38

0.33

FG%

0.43

0.37

0.47

0.41

0.44

0.39

FT%

0.72

0.72

0.76

0.71

0.77

0.76

TrueShooting%

0.50

0.48

0.55

0.49

0.54

0.50

Players from 2015-16 season with assigned clusters.

Player

Team

Cluster

10-Sean Clendinning

Algoma

1

24-Bacarius Dinkins

Lakehead

1

11-David Aromolaran

Laurentian

1

44-Sam Hirst

Laurentian

1

12-Matt Chesson

Laurier

1

03-Leon Alexander

McMaster

1

23-Aaron Redpath

McMaster

1

13-Marcus Lewis

Nipissing

1

21-Adika Peter-McNeilly

Ryerson

1

07-Ben Davis

Waterloo

1

10-Peter Scholtes

Western

1

22-Anthony Spiridis

Western

1

08-Nathan Culbreath

York

1

33-Matt Marshall

Brock

2

03-Nick Burke

Lakehead

2

21-Alexandre Robichaud

Lakehead

2

02-Simon Mikre

Laurier

2

04-Joey Puddister

Nipissing

2

06-Dylan Phillips

Waterloo

2

05-Tom Filgiano

Western

2

10-Mike Rocca

Windsor

2

13-Dani Elgadi

Brock

3

06-Devin Johnson

Toronto

3

12-Greg Morrow

Western

3

09-Alex Campbell

Windsor

3

05-Brett Zufelt

Algoma

4

06-Nathan Riley

Algoma

4

13-Reng Gum

Algoma

4

25-Tyler Brown

Brock

4

15-Drew Walford

Guelph

4

31-Jack Beatty

Guelph

4

10-Nick Simon

Laurentian

4

32-Joseph Sykes

Laurentian

4

03-Garrison Thomas

Laurier

4

04-Trevon McNeil

McMaster

4

21-David McCulloch

McMaster

4

22-Rohan Boney

McMaster

4

05-Marvin Ngonadi

Nipissing

4

07-Jerron Rhodes

Nipissing

4

09-Kalil Langston

Nipissing

4

01-Vikas Gill

Ottawa

4

05-Mehdi Tihani

Ottawa

4

09-Matt Plunkett

Ottawa

4

13-Nathan McCarthy

Ottawa

4

13-Sammy Ayisi

Queen's

4

20-Ryall Stroud

Queen's

4

06-Roshane Roberts

Ryerson

4

22-Juwon Grannum

Ryerson

4

05-Sage Usher

Toronto

4

09-Manny Sahota

Toronto

4

21-Daniel Johansson

Toronto

4

06-Alex Coote

Western

4

07-Jedson Tavernier

Western

4

15-Micah Kirubel

Windsor

4

22-Tyler Persaud

Windsor

4

04-Philip Gillen

York

4

05-Gene Spagnuolo

York

4

03-Andre Barber

Algoma

5

14-Ryan Bennett

Brock

5

03-Gavin Resch

Carleton

5

10-Connor Wood

Carleton

5
31-Guillaume Boucard
Carleton
5
41-Kaza Kajami-Keane
Carleton
5
04-Daniel Dooley
Guelph
5
05-Jonathan Wallace
Guelph
5
11-Taylor Boers
Guelph
5
05-Troy Joseph
McMaster
5
12-Tanner Graham
Queen's
5
04-Ammanuel Diressa
Ryerson
5
12-Aaron Best
Ryerson
5
04-Devon Williams
Toronto
5

11-Marko Kovac

Windsor

5

11-Tommy Hobbs

York

5

11-Johneil Simpson

Brock

6

05-Henry Tan

Lakehead

6

21-Anthony Iacoe

Laurentian

6

06-Will Coulthard

Laurier

6

11-Mike L'Africain

Ottawa

6

03-Jon Ravenhorst

Waterloo

6

13-Isiah Osborne

Windsor

6

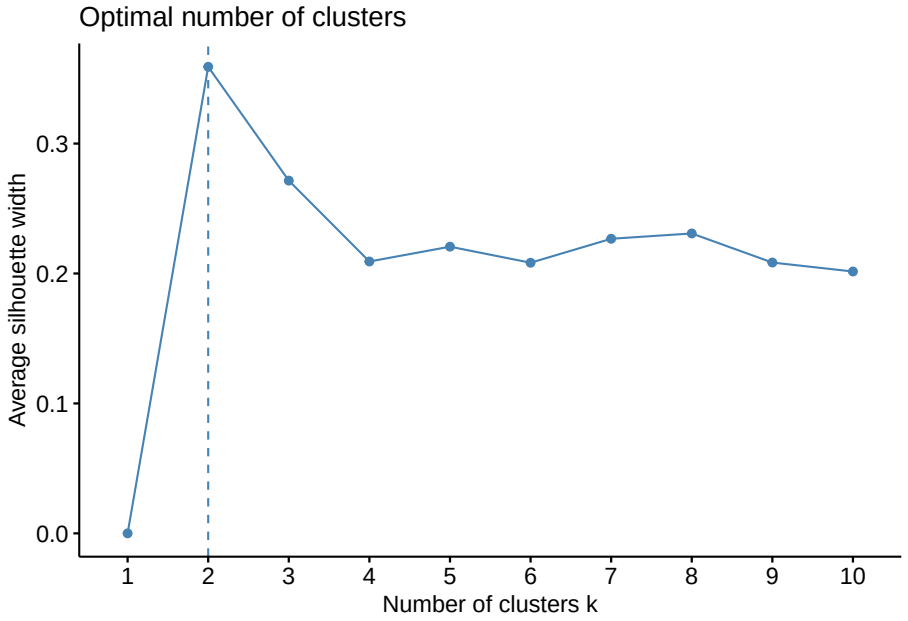


Figure 3.6: Silhouette Method for finding the optimal number of clusters for the 2016-17 season

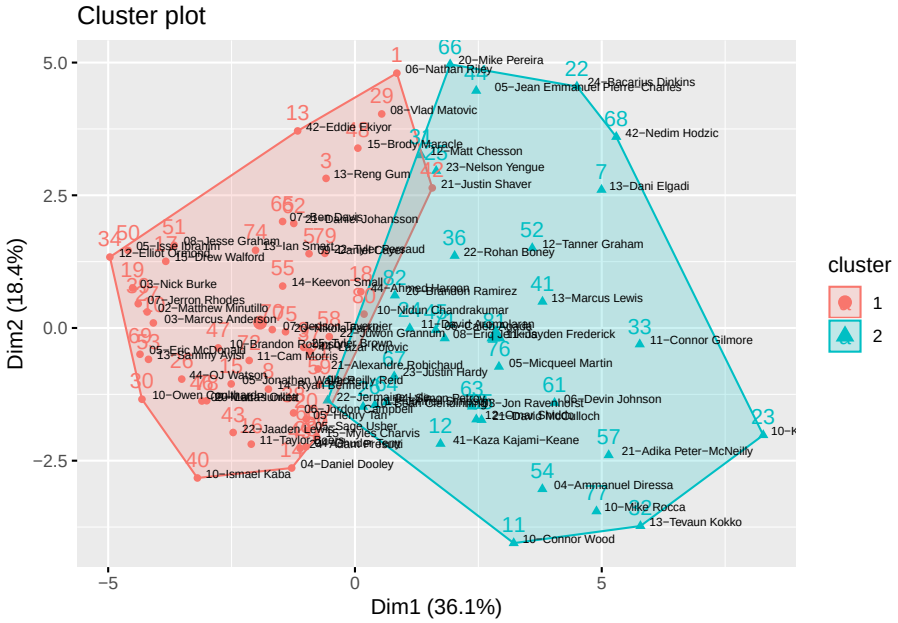


Figure 3.7: Cluster Plot of 2016-17 Season

3.4.3 2016-17 Season

The Silhouette method suggests 2 as the optimal number of clusters for this season. This is possibly separating the players into forwards/centers and guards. Below are tables to show the average statistics for each cluster and also which players belong to which cluster.

Average of each variable for each cluster for the 2016-17 season

Cluster 1

Cluster 2

3PointMade

16.67

25.09

3PointAttempted

51.80

73.85

Assists

30.02

42.82

Blocks

4.92

9.18

DefensiveRebounds

53.12

83.88

FieldGoalMade

55.86

101.97

FieldGoalAttempted

136.37

234.88

FreeThrowsMade

24.86

50.88

FreeThrowsAttempted

37.14

71.27

Minutes

452.90

559.09

OffensiveRebounds

20.53

30.91

PersonalFouls

35.53

44.12

Points

153.24

279.91

Rebounds

73.65

114.79

Steals

16.88

23.67

Turnovers

28.86

47.39

Home

9.14

9.70

GamesPlayed

18.43

19.12

PointsPerGame

8.34

14.70

MinutesPerGame

24.59

29.24

3P%

0.28

0.30

FG%

0.41

0.44

FT%

0.68

0.71

TrueShooting%

0.50

0.52

Players from 2016-17 season with assigned clusters.

Player

Team

Cluster

06-Nathan Riley

Algoma

1

13-Reng Gum

Algoma

1

09-Daniel Cayer

Brock

1

14-Ryan Bennett

Brock

1

25-Tyler Brown

Brock

1

03-Marcus Anderson

Carleton

1

42-Eddie Ekiyor

Carleton

1

04-Daniel Dooley

Guelph

1

05-Jonathan Wallace

Guelph

1

11-Taylor Boers

Guelph

1

15-Drew Walford

Guelph

1

44-Ahmed Haroon

Guelph

1

03-Nick Burke

Lakehead

1

05-Henry Tan

Lakehead

1

21-Alexandre Robichaud

Lakehead

1

44-OJ Watson

Laurentian

1

02-Matthew Minutillo

Laurier

1

04-Chuder Teny

Laurier

1

08-Vlad Matovic

Laurier

1

10-Owen Coulthard

Laurier

1

12-Elliot Ormond

McMaster

1

44-Lazar Kojovic

McMaster

1

06-Jordon Campbell

Nipissing

1

07-Jerron Rhodes

Nipissing

1

10-Ismael Kaba

Nipissing

1

21-Justin Shaver

Nipissing

1

22-Jaaden Lewis

Nipissing

1

09-Matt Plunkett

Ottawa

1

10-Brandon Robinson

Ottawa

1

15-Brody Maracle

Ottawa

1

24-Adam Presutti

Ottawa

1

05-Isse Ibrahim

Queen's

1

08-Jesse Graham

Queen's

1

13-Sammy Ayisi

Queen's

1

14-Keevon Small

Ryerson

1

15-Myles Charvis

Ryerson

1

22-Juwon Grannum

Ryerson

1

04-Reilly Reid

Toronto

1

05-Sage Usher

Toronto

1

21-Daniel Johansson

Toronto

1

07-Ben Davis

Waterloo

1

05-Eric McDonald

Western

1

07-Jedson Tavernier

Western

1

11-Cam Morris

Western

1

13-Ian Smart

Western

1

20-Nikola Farkic

Western

1

20-Lucas Orlita

Windsor

1

22-Tyler Persaud

Windsor

1

10-Nidun Chandrakumar

York

1

10-Sean Clendinning

Algoma

2

22-Jermaine Lyle

Algoma

2

11-Johneil Simpson

Brock

2

13-Dani Elgadi

Brock

2

10-Connor Wood

Carleton

2

41-Kaza Kajami-Keane

Carleton

2

24-Bacarius Dinkins

Lakehead

2

10-Kadre Gray

Laurentian

2

11-David Aromolaran

Laurentian

2

23-Nelson Yengue

Laurentian

2

12-Matt Chesson

Laurier

2

13-Tevaun Kokko

Laurier

2

11-Connor Gilmore

McMaster

2

21-David McCulloch

McMaster

2

22-Rohan Boney

McMaster

2

13-Marcus Lewis

Nipissing

2

05-Jean Emmanuel Pierre-Charles

Ottawa

2

06-Caleb Agada

Ottawa

2

12-Tanner Graham

Queen's

2

04-Ammanuel Diressa

Ryerson

2

21-Adika Peter-McNeilly

Ryerson

2

06-Devin Johnson

Toronto

2

03-Jon Ravenhorst

Waterloo

2

04-Simon Petrov

Waterloo

2

20-Mike Pereira

Waterloo

2

23-Justin Hardy

Waterloo

2

42-Nedim Hodzic

Waterloo

2

08-Eriq Jenkins

Western

2

12-Omar Shiddo

Western

2

05-Micqueel Martin

Windsor

2

10-Mike Rocca

Windsor

2

11-Jayden Frederick

York

2

20-Brandon Ramirez

York

2

3.4.4 2017-18 Season

The optimal number of clusters suggested by the D-index method is 5 for this season. This is possibly separating the players into the actual posistions (Point Guard, Shooting Guard, Small Forward, Power Forward, and Center). Below are tables to show the average statistics for each cluster and also which players belong to which cluster.

Average of each variable for each cluster for the 2017-18 season

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

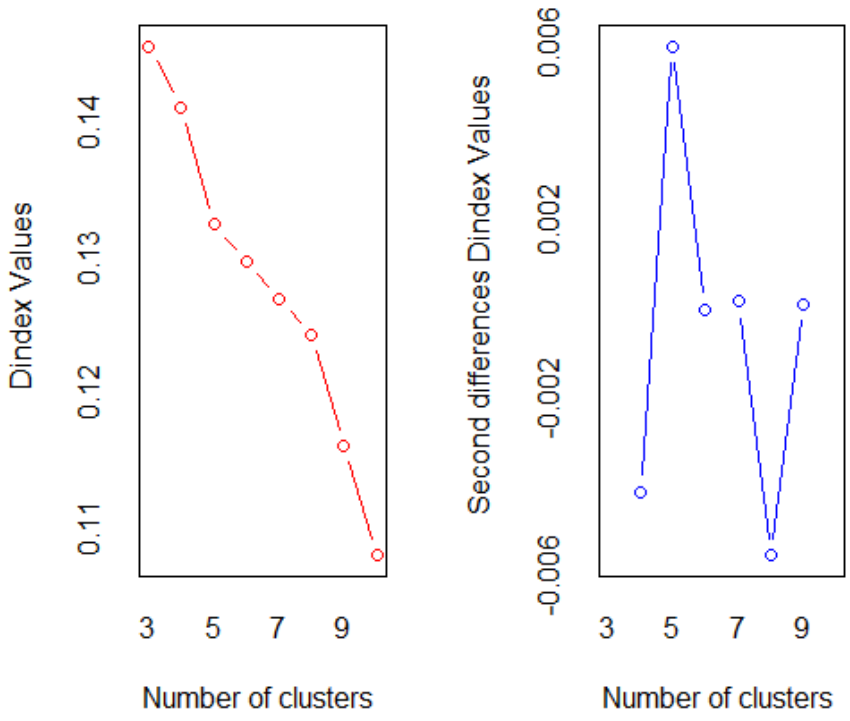


Figure 3.8: Finding the Optimal Cluster using D-index for the 2017-18 season.

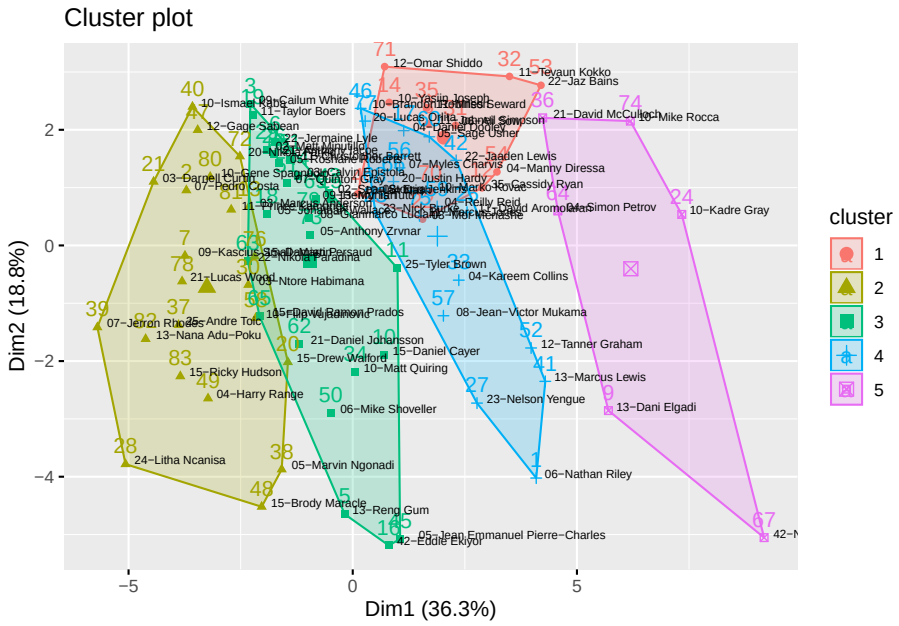


Figure 3.9: Cluster Plot of 2017-18 Season

3PointMade

46.75

16.23

23.72

28.11

29.50

3PointAttempted

127.75

49.32

71.44

85.00

86.33

Assists

57.92

33.95

41.36

64.17

85.33

Blocks

3.83

6.68

10.04

11.06

11.33

DefensiveRebounds

70.17

60.45

74.40

93.22

137.00

FieldGoalMade

133.00

53.00

81.40

109.00

156.33

FieldGoalAttempted

309.83

129.64

189.52

255.67

340.33

FreeThrowsMade

58.42

20.86

36.56

55.17

117.17

FreeThrowsAttempted

75.50

33.68

52.52

77.28

144.00

Minutes

606.67

496.41

571.40

703.78

750.17

OffensiveRebounds

22.17

23.73

26.32

30.72

40.17

PersonalFouls

49.33

44.00

47.08

52.89

54.00

Points

371.17

143.09

223.08

301.28

459.33

Rebounds

92.33

84.18

100.72

123.94

177.17

Steals

26.50

17.73

19.88

31.11

27.67

Turnovers

48.58

31.09

34.00

49.33

65.33

Home

11.25

11.36

11.44

11.56

11.50

GamesPlayed

22.25

22.50

22.72

23.22

23.17

PointsPerGame

16.82

6.44

9.88

13.00

19.95

MinutesPerGame

27.36

22.11

25.22

30.34

32.43

3P%

0.37

0.27

0.29

0.30

0.27

FG%

0.43

0.41

0.43

0.43

0.46

FT%

0.77

0.64

0.71

0.71

0.82

TrueShooting%

0.54

0.48

0.52

0.52

0.56

Players from 2017-18 season with assigned clusters.

Player

Team

Cluster

10-Ian Nash

Algoma

1

11-Johneil Simpson

Brock

1

35-Cassidy Ryan

Brock

1

10-Yasiin Joseph

Carleton

1

08-Mor Menashe

Lakehead

1

06-Ali Sow

Laurier

1

11-Tevaun Kokko

Laurier

1

11-Miles Seward

McMaster

1

22-Jaz Bains

Queen's

1

04-Manny Diressa

Ryerson

1

10-Marko Kovac

Western

1

12-Omar Shiddo

Western

1

07-Pedro Costa

Algoma

2

09-Kascius Small-Martin

Brock

2

03-Marcus Anderson

Carleton

2

15-Drew Walford

Guelph

2

03-Darnell Curtin

Lakehead

2

24-Litha Ncanisa

Laurentian

2

03-Ntore Habimana

Laurier

2

25-Andre Toic

McMaster

2

05-Marvin Ngonadi

Nipissing

2

07-Jerron Rhodes

Nipissing

2

10-Ismael Kaba

Nipissing

2

12-Gage Sabean

Ottawa

2

15-Brody Maracle

Ottawa

2

04-Harry Range

Queen's

2

10-Filip Vujadinovic

Ryerson

2

20-Nikola Farkic

Western

2

15-Damian Persaud

Windsor

2

21-Lucas Wood

Windsor

2

10-Gene Spagnuolo

York

2

11-Prince Kamunga

York

2

13-Nana Adu-Poku

York

2

15-Ricky Hudson

York

2

09-Cailum White

Algoma

3

13-Reng Gum

Algoma

3

22-Jermaine Lyle

Algoma

3

15-Daniel Cayer

Brock

3

25-Tyler Brown

Brock

3

13-Munis Tutu

Carleton

3

42-Eddie Ekiyor

Carleton

3

05-Jonathan Wallace

Guelph

3

11-Taylor Boers

Guelph

3

21-Anthony Iacoe

Laurentian

3

02-Matt Minutillo

Laurier

3

10-Matt Quiring

McMaster

3

02-Sean Stoqua

Ottawa

3

03-Calvin Epistola

Ottawa

3

05-Jean Emmanuel Pierre-Charles

Ottawa

3

06-Mike Shoveller

Queen's

3

07-Quinton Gray

Queen's

3

05-Roshane Roberts

Ryerson

3

11-Christopher Barrett

Toronto

3

21-Daniel Johansson

Toronto

3

22-Nikola Paradina

Toronto

3

15-David Ramon Prados

Waterloo

3

09-Henry Tan

Western

3

05-Anthony Zrvnar

Windsor

3

08-Gianmarco Luciani

York

3

06-Nathan Riley

Algoma

4

04-Daniel Dooley

Guelph

4

23-Nick Burke

Lakehead

4

11-David Aromolaran

Laurentian

4

23-Nelson Yengue

Laurentian

4

04-Kareem Collins

McMaster

4

13-Marcus Lewis

Nipissing

4

22-Jaaden Lewis

Nipissing

4

10-Brandon Robinson

Ottawa

4

12-Tanner Graham

Queen's

4

07-Myles Charvis

Ryerson

4

08-Jean-Victor Mukama

Ryerson

4

04-Reilly Reid

Toronto

4

05-Sage Usher

Toronto

4

20-Justin Hardy

Waterloo

4

08-Eriq Jenkins

Western

4

11-Marcus Jones

Windsor

4

20-Lucas Orlita

Windsor

4

13-Dani Elgadi

Brock

5

10-Kadre Gray

Laurentian

5

21-David McCulloch

McMaster

5

04-Simon Petrov

Waterloo

5

42-Nedim Hodzic

Waterloo

5

10-Mike Rocca

Windsor

5

3.4.5 2018-19 Season

The optimal number of clusters suggested for the 2018-19 season is 4. Below are tables to show the average statistics for each cluster and also which players belong to which cluster.

Average of each variable for each cluster for the 2018-19 season

Cluster 1

Cluster 2

Cluster 3

Cluster 4

3PointMade

25.24

75.50

18.74

38.00

3PointAttempted

79.55

205.75

58.26

106.56

Assists

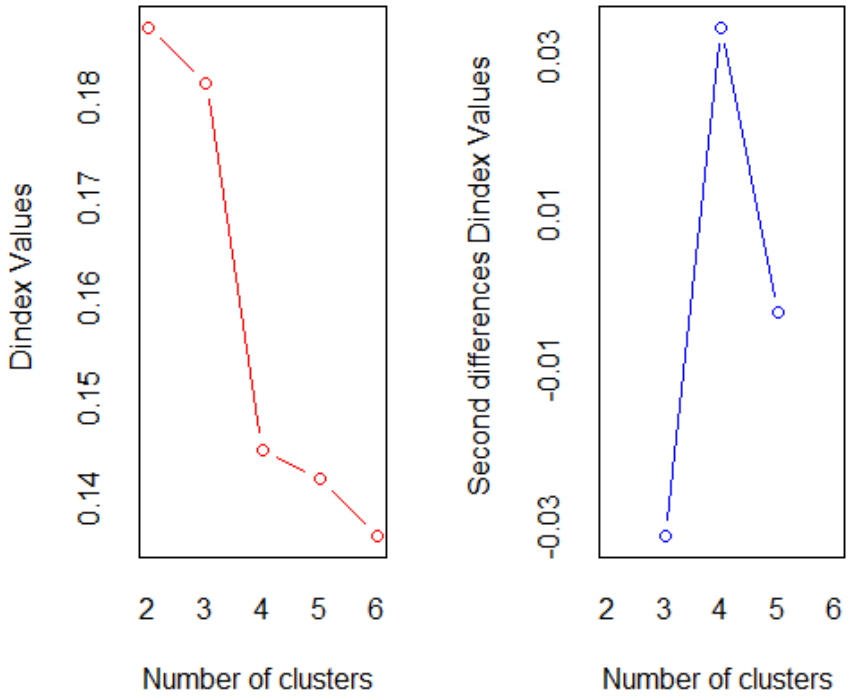


Figure 3.10: Finding the Optimal Cluster using D-index for the 2018-19 season.

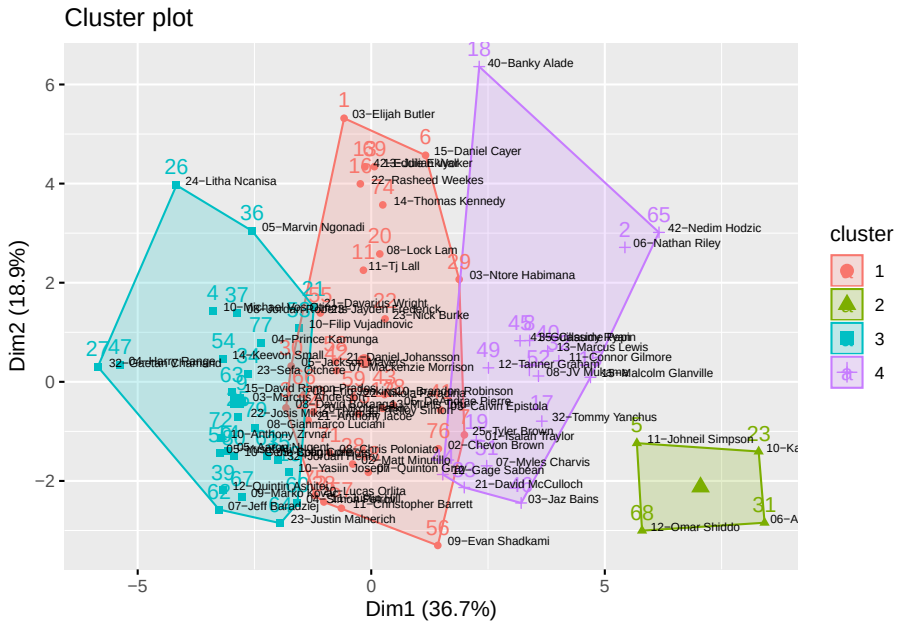


Figure 3.11: Cluster Plot of 2018-19 Season

51.12

70.75

29.81

53.06

Blocks

9.03

4.25

6.85

11.12

DefensiveRebounds

83.12

96.00

60.33

114.81

FieldGoalMade

91.48

192.75

62.52

132.25

FieldGoalAttempted

215.67

435.75

150.70

308.19

FreeThrowsMade

40.79

105.00

28.93

81.19

FreeThrowsAttempted

58.48

130.75

40.74

108.38

Minutes

613.58

775.25

478.74

708.94

OffensiveRebounds

28.67

22.00

20.56

36.44

PersonalFouls

52.00

48.50

40.37

50.38

Points

249.00

566.00

172.70

383.69

Rebounds

111.79

118.00

80.89

151.25

Steals

24.64

31.00

15.96

24.69

Turnovers

37.42

68.00

29.00

50.12

Home

11.55

11.50

10.33

11.38

GamesPlayed

22.73

23.00

20.52

22.88

PointsPerGame

11.00

24.88

8.65

16.80

MinutesPerGame

27.05

33.72

23.51

30.99

3P%

0.29

0.36

0.29

0.33

FG%

0.43

0.44

0.42

0.43

FT%

0.69

0.79

0.70

0.75

TrueShooting%

0.51

0.57

0.51

0.54

Players from 2018-19 season with assigned clusters.

Player

Team

Cluster

03-Elijah Butler

Algoma

1

08-David Bokanga

Algoma

1

15-Daniel Cayer

Brock

1

25-Tyler Brown

Brock

1

11-Tj Lall

Carleton

1

13-Munis Tutu

Carleton

1

42-Eddie Ekiyor

Carleton

1

22-Rasheed Weekes

Guelph

1

08-Lock Lam

Lakehead

1
23-Nick Burke
Lakehead
1
21-Anthony Iacoe
Laurentian
1
02-Matt Minutillo
Laurier
1
03-Ntore Habimana
Laurier
1
05-Jackson Mayers
Laurier
1
11-Justin Hill
Nipissing
1
03-Calvin Epistola
Ottawa
1
07-Mackenzie Morrison
Ottawa
1
10-Brandon Robinson
Ottawa
1
07-Quinton Gray
Queen's
1

23-Jayden Frederick

Ryerson

1

09-Evan Shadkami

Toronto

1

11-Christopher Barrett

Toronto

1

21-Daniel Johansson

Toronto

1

22-Nikola Paradina

Toronto

1

08-Eriq Jenkins

Western

1

13-Julian Walker

Western

1

20-Nikola Farkic

Western

1

08-Chris Poloniato

Windsor

1

11-Telloy Simon

Windsor

1

14-Thomas Kennedy

Windsor

1

20-Lucas Orlita

Windsor

1

02-Chevon Brown

York

1

05-DeAndrae Pierre

York

1

11-Johneil Simpson

Brock

2

10-Kadre Gray

Laurentian

2

06-Ali Sow

Laurier

2

12-Omar Shiddo

Western

2

10-Michael Vos Otin

Brock

3

03-Marcus Anderson

Carleton

3

10-Yasiin Joseph

Carleton

3

05-Aaron Nugent

Guelph

3

21-Davarius Wright

Lakehead

3

22-Josis Mikia-Thomas

Laurentian

3

24-Litha Ncanisa

Laurentian

3

32-Gaetan Chamand

Laurentian

3

23-Sefa Otchere

McMaster

3

32-Jordan Henry

McMaster

3

05-Marvin Ngonadi

Nipissing

3

08-Jordan Roberts

Nipissing

3

12-Quintin Ashitei

Nipissing

3

04-Harry Range

Queen's

3

05-Yusuf Ali

Ryerson

3

10-Filip Vujadinovic

Ryerson

3

14-Keevon Small

Ryerson

3

04-Simon Petrov

Waterloo

3

05-Colin Connors

Waterloo

3

07-Jeff Baradziej

Waterloo

3

15-David Ramon Prados

Waterloo

3

23-Justin Malnerich

Waterloo

3

09-Marko Kovac

Western

3

10-Anthony Zrvnar

Windsor

3

04-Prince Kamunga

York

3

08-Gianmarco Luciani

York

3

10-Gene Spagnuolo

York

3

06-Nathan Riley

Algoma

4

35-Cassidy Ryan

Brock

4

15-Malcolm Glanville

Guelph

4

32-Tommy Yanchus

Guelph

4

40-Banky Alade

Guelph

4

01-Isaiah Traylor

Lakehead

4

11-Connor Gilmore

McMaster

4

21-David McCulloch

McMaster

4

13-Marcus Lewis

Nipissing

4

12-Gage Sabean

Ottawa

4

41-Guillaume Pepin

Ottawa

4

03-Jaz Bains

Queen's

4

12-Tanner Graham

Queen's

4

07-Myles Charvis

Ryerson

4

08-JV Mukama

Ryerson

4

42-Nedim Hodzic

Waterloo

4

3.5 Conclusion

Unsupervised learning used on basketball data can be very helpful. It can be used to categorize players and to see what their style of play is. It can also be used for match-ups and for predicting important players. For instance, if you find that a player was in the same cluster as the catch and shoot players, a coach can assign an appropriate defender. Knowing the style of play for your opponents is very useful for defensive purposes. In my opinion the higher the number of clusters assigned, the better because it would distinguish the type of player more.

Chapter 4

Classification of Wins

In machine learning and statistics, classification is a supervised learning approach in which the machine learns from the data input given to it and then uses this learning to classify new observations. In this case, classification can be used to identify a win and loss and also to predict whether a game will be a win or loss. That means we want to identify which variables are the most important in distinguishing a win (or a loss_. There are many types of classification techniques such as Random Forests, Support Vector Machines, Logistic Regression, XGBoost, etc..

4.1 Random Forests

Random forest is an ensemble learning method for classification. Ensemble methods are very effective because they use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the learning algorithms alone. A random forest consists of a large number of decision trees that operate as an ensemble. Each individual tree in the random forest gives a prediction of outcome and the class with the most votes becomes the model's prediction[11]. The reason why a random forest is a great technique is because a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. Random forests also give an importance score for all the features used in the model. A standard procedure is to first use all the variables and then use feature importance to narrow the model down to get more accurate results.

4.1.1 Model 1

In this random forest model we are predicting wins using the following predictors: Assists,DefensiveRebounds,TotalRebounds,Turnovers,PushBallfromTurnover,Steals,

PressOffense,UnguardedJumpShots,AllFreeThrows,P&RBallHandler-SingleCovered,Cuts, GuardedJumpShots,ShortJumpShots,TransitionOffense,LongJumpShots,Transitions,SpotUps, P&RBallHandler-DefenseCommits,PushBallfromShotAttempt,PushBalltoHalfCourtOff., OffensiveRebounds,MiscellaneousPossessions,Isolation-SingleCovered,Post-Up-SingleCovered, MediumJumpShots,Blocks,OffScreens,Off.Reb.-PutBacks,Handoffs,Off.Reb.-ResetOffense, TransitionTurnover,P&RRollMan,Isolation-DefenseCommits,Post-Up-DefenseCommits, Post-Up-HardDoubleTeam,P&RBallHandler-Traps. The model is trained on a train set which is a random sample (without replacement) of 70% of the dataset and tested on a random sample of 30% of the dataset. The accuracy score is obtained below.

```
Accuracy: 0.7477064220183486
```

A very important perk of the random forest algorithm is it allows us to obtain the Feature importance to let us know which variables were the most important for creating the model, i.e. which features are the most important in classifying and predicting wins. A table of the Feature importance from this model is shown below.

Table 4.1: Feature Importance of Random Forest Model 1

Feature	Feature Importance Value
Assists	0.096881
DefensiveRebounds	0.062817
TotalRebounds	0.059851
Turnovers	0.056744
PushBallfromTurnover	0.044454
Steals	0.038865
PressOffense	0.037174
Unguarded Jump Shots	0.030118
AllFreeThrows	0.029086
P&RBallHandler-SingleCovered	0.028485
Cuts	0.027246
GuardedJumpShots	0.026854
ShortJumpShots	0.025915
TransitionOffense	0.025691
LongJumpShots	0.023443
Transitions	0.023180
SpotUps	0.023102
P&RBallHandler-DefenseCommits	0.022575
PushBallfromShotAttempt	0.022505
PushBalltoHalfCourtOff.	0.022279
OffensiveRebounds	0.022109
MiscellaneousPossessions	0.020995
Isolation-SingleCovered	0.020878

Feature	Feature Importance Value
Post-Up-SingleCovered	0.020679
MediumJumpShots	0.019900
Blocks	0.019759
OffScreens	0.018904
Off.Reb.-PutBacks	0.017528
Handoffs	0.017504
Off.Reb.-ResetOffense	0.017282
TransitionTurnover	0.017150
P&RRollMan	0.015011
Isolation-DefenseCommits	0.013565
Post-Up-DefenseCommits	0.013291
Post-Up-HardDoubleTeam	0.012348
P&RBallHandler-Traps	0.005832

Note: since random forests take samples randomly, the accuracy will vary depending on the seed chosen.

4.1.2 Model 2

In this random forest model we are predicting wins using a refined selection of predictors: Assists,DefensiveRebounds,TotalRebounds,Turnovers,Steals,PushBallfromTurnover,PressOffense,AllFreeT Transitions,GuardedJumpShots,PushBallfromShotAttempt,AllP&RBallHandler,SpotUps, AllPost-Up,PushBalltoHalfCourtOff.,AllOffensiveRebounds,MiscellaneousPossessions, AllIsolation,OffScreens,Blocks,MediumJumpShots,Isolation-SingleCovered,Handoffs.The model is trained on a train set which is a random sample (without replacement) of 70% of the dataset and tested on a random sample of 30% of the dataset. The accuracy score is obtained below.

```
Accuracy: 0.7477064220183486
```

A table of the Feature importance from this model is shown below

Table 4.2: Feature Importance of Random Forest Model 2

Feature	Feature Importance Value
Assists	**0.103865**
DefensiveRebounds	0.079669
TotalRebounds	0.064047
Turnovers	0.062428
Steals	0.059118
PushBallfromTurnover	0.048081

Feature	Feature Importance Value
PressOffense	0.039639
AllFreeThrows	0.034666
Unguarded Jump Shots	0.033493
ShortJumpShots	0.032340
Cuts	0.031976
LongJumpShots	0.030290
Transitions	0.029443
GuardedJumpShots	0.029287
PushBallfromShotAttempt	0.028180
AllP&RBallHandler	0.028030
SpotUps	0.027917
AllPost-Up	0.027047
PushBalltoHalfCourtOff.	0.025952
AllOffensiveRebounds	0.025533
MiscellaneousPossessions	0.024643
AllIsolation	0.024239
OffScreens	0.023363
Blocks	0.022834
MediumJumpShots	0.022212
Isolation-SingleCovered	0.021883
Handoffs	0.019824

Assists are the most important feature in both models for classifying whether a game is a win or loss.

4.2 Logistic Regression

Logistic Regression is a form of regression that is used when the response variable is a categorical variable [12]. In this case it is a binary value (e.g. Success or Failure). The game by game data can be used to create a model that predicts Wins.

4.2.1 Model

The same features are used in this model as the second model in the Random Forests section. Logistic regression will be used to classify and then predict wins. The equation is below

$$Win = \beta_0 + \beta_1 * Assists + ... + \beta_{27} * Handoffs$$

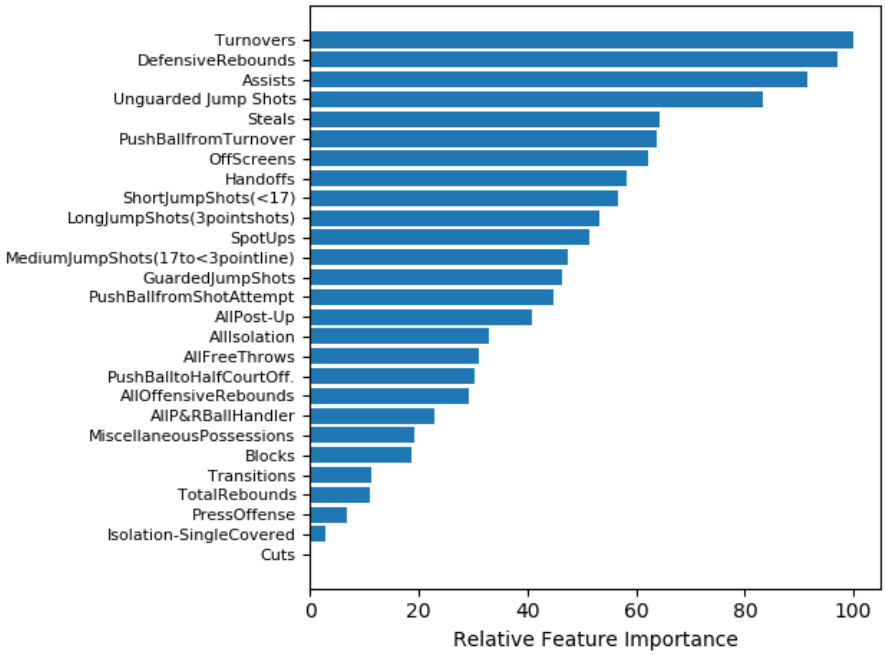


Figure 4.1: Feature Importance from Logistic Regression Model.

Again, the model is trained on a train set which is a random sample (without replacement) of 70% of the dataset and tested on a random sample of 30% of the dataset. The accuracy score is obtained below.

```
Accuracy: 0.7821100917431193
```

The feature importance from Logistic Regression differs from Random Forests.Although, Defensive Rebounds, Assists, and Turnovers are still on the top of the list. In general, turnovers negatively impact teams and can be an important feature to distinguish teams that are less likely to win if they make more turnovers.

4.2.2 Assists

A dataset has been modified to subtract the home team's statistics from the away team's statistics for each game so that there are differential statistics. The differential statistics were compared to see which contributed to the highest proportion of wins.

Table 4.3: Differential Statistics & Proportion of Wins

Differential Statistics	Proportion of Wins
Positive Assists Differential	845/1171 = 72.2%
Positive Rebounds Differential	817/1171 = 69.8%
Negative Turnovers Differential	726/1171 = 62%

4.2.2.1 Risk Ratio & Odds Ratio

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Win
Comparing : Positive Assists Differential vs. Negative Assists Differential

                              Win Lose    P(Win) 95% conf. interval
Positive Assists Differential 845  326    0.7216    0.6952   0.7465
Negative Assists Differential 326  845    0.2784    0.2535   0.3048

                                   95% conf. interval
            Relative Risk: 2.5920    2.3481   2.8613
         Sample Odds Ratio: 6.7186    5.6078   8.0494
Conditional MLE Odds Ratio: 6.7123    5.5848   8.0849
    Probability difference: 0.4432    0.4059   0.4784

             Exact P-value: 0.0000
        Asymptotic P-value: 0.0000
------------------------------------------------------
```

Above is a two-by-two table analysis. The Sample Odds Ratio tells us that odds of a team winning is 6.7 higher given they have more assists than their opponent compared to teams that have fewer assists than their opponent. The Relative Risk tells us that teams with more assists than their opponent have 2.59 times the 'risk' of winning compared to teams with fewer assists than their opponent.

4.2.3 Why are Assists so important?

Assists can lead to effective scoring. A player is getting set up for a shot and each team can distribute their shots differently. A study was done in the NBA (Pelechrinis, Konstantinos, 2019) [13] that has shown that on average an assisted shot added 0.16 expected points more compared to an unassisted shot. If teams looked for the extra pass on 15 of their unassisted shots, this corresponds to approximately 2.4 additional expected points over the course of the game.

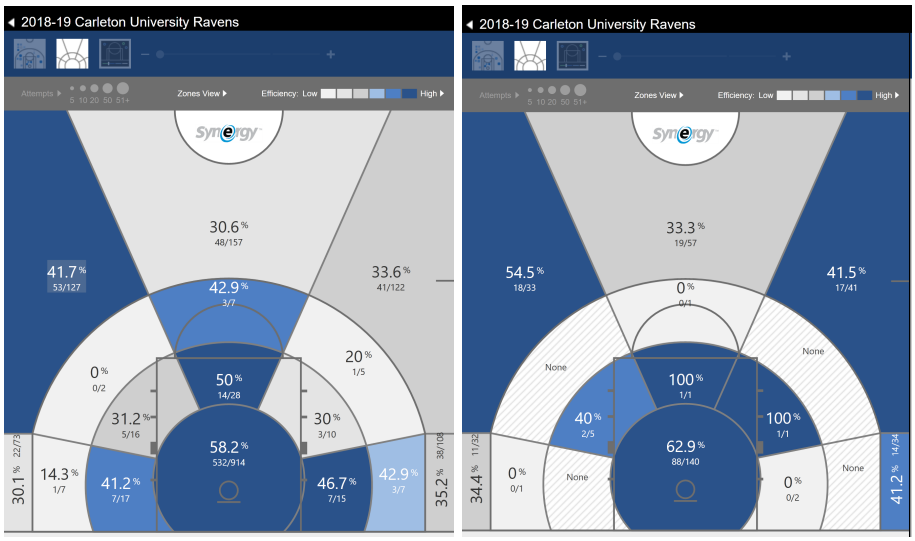


Figure 4.2: Side-by-Side shot chart of Carleton's 2018-19 season. The left side is the shot chart of the entire season without any filters. The right side shows the chart of the entire season where shots were derived from passing plays.

An assist can increase the average field goal percentage of a type of shot as opposed to an unassisted shot (Pelechrinis, Konstantinos, 2019). Also, assists are necessary for effective play making. As seen previously, transitions, spot-ups and cuts are all very effective offensive plays and the thing that connects them together is an assist.

4.2.3.1 Shots Derived From Assists

Using Synergy's Multi-Game Shot Chart it is possible to see the difference in shooting efficiency between shots derived from an assist and shots that were not.

Chapter 5

Final Words

We have finished a nice book.