NOVA UNIVERSITY
LISBON

The Dynamics of Basketball Performance

Author: **Mattia Piccinato, Matteo Saterini**

Academic Year: 2023-24

Contents

1 | Introduction

1.1. Overview

In a world where data are more abundant and available than ever before, sports analytics has emerged as a really strong tool that can help teams and organizations in decision-making and in understanding athletic performance with the goal of optimizing it.

The National Collegiate Athletic Association (NCAA) basketball tournaments entertain millions of fans annually, displaying the raw talent and strategic process of collegiate athletes, with a business ecosystem that drives billions of dollars into the American sports industry; the choice of analyzing a basketball dataset becomes then a great example of what a professional data scientist working for an NCAA team could do in his workdays.

Moreover, in an era characterized by evolving gameplay tactics and technological advancements, unveiling hidden patterns might be the best way to maximize results: we aim at finding and explaining strategic nuances that can inform coaching decisions, player development programs, and tactical maneuvers.

In this report, we embark on a journey into the heart of college basketball analytics, driven by the conviction that data has the power to unlock new dimensions of understanding and transform the game.

1.2. Structure of the dataset

The dataset contains statistics from the college basketball season from the year 2015 to 2019. Each row contains the statistics of every team that played in the NCAA during a given season.

The whole dataset has 1754 rows and 24 columns, and the features are the following:

- G: Number of games played

- W: Number of games won
- ADJOE: Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense)
- ADJDE: Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division I offense)
- BARTHAG: Power Rating (Chance of beating an average Division I team)
- EFGO: Effective Field Goal Percentage Shot
- EFGD: Effective Field Goal Percentage Allowed
- TOR: Turnover Percentage Allowed (Turnover Rate)
- TORD: Turnover Percentage Committed (Steal Rate)
- ORB: Offensive Rebound Rate
- DRB: Offensive Rebound Rate Allowed
- FTR : Free Throw Rate (How often the given team shoots Free Throws)
- FTRD: Free Throw Rate Allowed
- 2PO: Two-Point Shooting Percentage
- 2PD: Two-Point Shooting Percentage Allowed
- 3PO: Three-Point Shooting Percentage
- 3PD: Three-Point Shooting Percentage Allowed
- ADJT: Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)
- WAB: Wins Above Bubble (The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it)
- POSTSEASON: Round where the team's season ended (R68 = First Four, R64 = Round of 64, R32 = Round of 32, S16 = Sweet Sixteen, E8 = Elite Eight, F4 = Final Four, 2ND = Runner-up, Champion = Winner of the NCAA March Madness Tournament)
- SEED: Seed in the NCAA March Madness Tournament
- YEAR: Season

For our analysis, we will only consider the data belonging to year 2015. Indeed, by focusing on a single season, you ensure consistency in the data collection methodologies, rules, and regulations governing the games. Moreover, it's necessary to only consider belonging to a specific year, in order to have a indipendency between observations, since teams statistics are correlated and dependent throughout the years.

This consistency enhances the comparability of the data and reduces potential confounding variables that might arise from changes in rules or data collection practices across different seasons. Thus, overall, we will only take into consideration the 351 entities belonging to year 2015.

1.3. Scope

We will delve into a comprehensive analysis of the NCAA basketball dataset utilizing various data analysis methods and algorithms. We will primarily focus on two major methodologies: Regression Analysis and Principal Component Analysis (PCA).

Within Regression Analysis, we will explore the relationship between select features through linear regression, investigating normality, correlation coefficients, hypothesis testing, and confidence intervals.

And, following that, we will employ Principal Component Analysis to visualize and interpret patterns within a subset of features related to specific aspects of basketball performance.

Furthermore, we will venture into Fuzzy Clustering, specifically studying the Fuzzy C-Means (FCM) program. In order to identify the best number of clusters we will explore various techniques, comprehending the Iterative Anomalous Pattern (IAP) clustering algorithm. We will analyze the results obtained evaluating the effectiveness in uncovering meaningful insights from the dataset.

1.4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) serves as a crucial initial step in the data mining process, offering insights into the underlying patterns, trends, and relationships within the dataset. By employing a combination of statistical techniques, visualization methods, and domain knowledge, EDA facilitates a deeper understanding of the data's structure and characteristics.

In this section, we approach a comprehensive exploration of the dataset at hand. We aim to

uncover key features, detect potential outliers or anomalies, identify any missing or incorrect values, and evaluate the distributional properties of the variables. Through graphical representations and summary statistics, we seek to elucidate the inherent properties of the data, building the way for subsequent modeling and analysis tasks.

As a preliminary step, we create a variable that will be highly useful throughout this analysis: the "win ratio", calculated as the number of matches won (W) divided by the number of matches played (G). This variable cleverly addresses the issue that not all teams play the same number of matches, thereby mitigating bias that would arise from studying only the W variable.

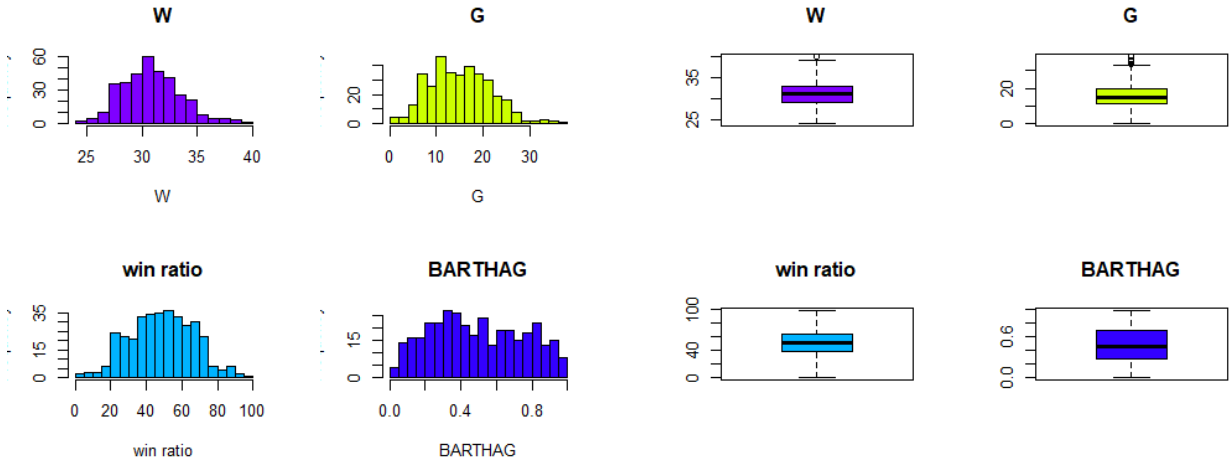
Let's now have a look at the summary statistics of the features that we are going to consider in the statistical analysis:

"	""Min.""	""1st Qu.""	""Median""	""Mean""	""3rd Qu.""	""Max.""
G	24	29	31	31.33	33	40
W	0	11	15	15.67	20	38
win_ratio	0	37.08	50	49.02	62.33	97.44
BARTHAG	0.0077	0.2867	0.4677	0.4939	0.704	0.9842
ADJOE	76.7	97.9	102.1	102.3	106.5	129.1
ADJDE	84	97.75	102.6	102.31	107.1	120
EFG_O	39.4	46.9	49	48.97	50.8	58.3
EFG_D	39.6	47.05	49.2	49.18	51.25	55.8
TOR	12.4	17.8	19.2	19.13	20.3	26.1
TORD	14	17.6	18.9	19.04	20.4	28
ORB	19.3	28.65	30.8	30.86	33.7	42.1
DRB	22.4	29.2	31.2	31.09	32.9	40.4
FTR	26.1	33.55	36.6	37.04	40.3	51
FTRD	22.4	33.1	36.6	37.33	41.5	55.5
ADJ_T	57.2	63.3	64.9	64.94	66.6	77.3
WAB	-23.2	-12.8	-8.4	-7.765	-3.3	13.1
X2P_O	38.4	45.6	47.6	47.73	49.7	58.2
X2P_D	37.7	45.6	48	47.92	50.15	56.5
X3P_O	25.2	32.1	34.2	34.17	36.05	44.1
X3P_D	27.1	32.7	34.3	34.4	36.15	40.3

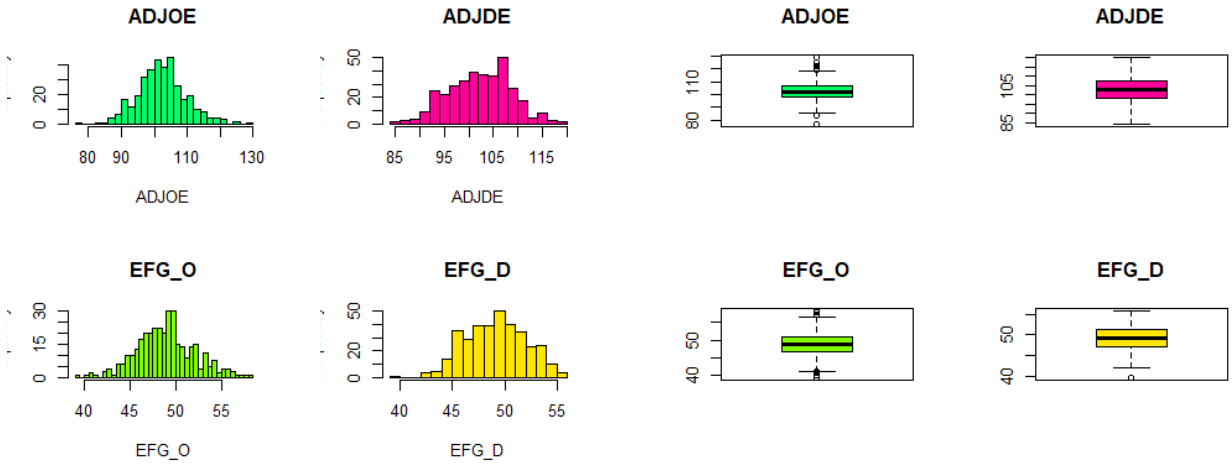
Table 1.1: Summary Table

Let's also have a look at the plot of the distributions through histograms and boxplots:

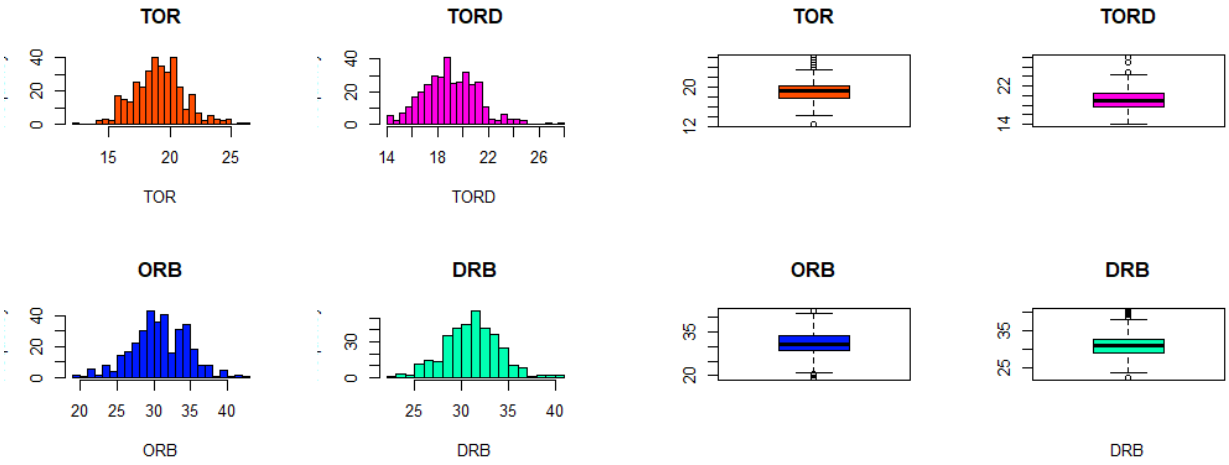
- These four features are the most general ones, and talk about how much a team played and won.



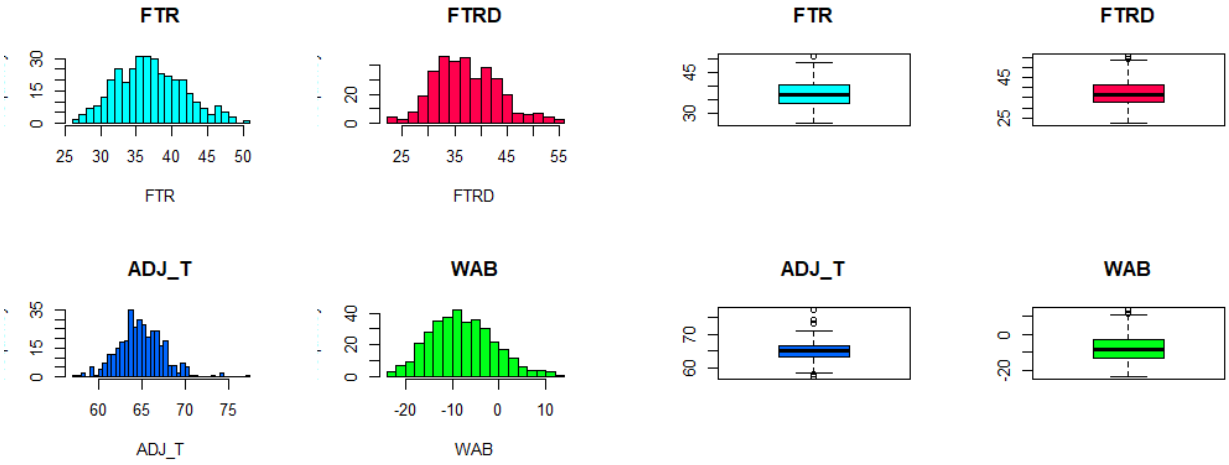
- These four features all talk about the efficiency of a team, both on the offensive and defensive side.



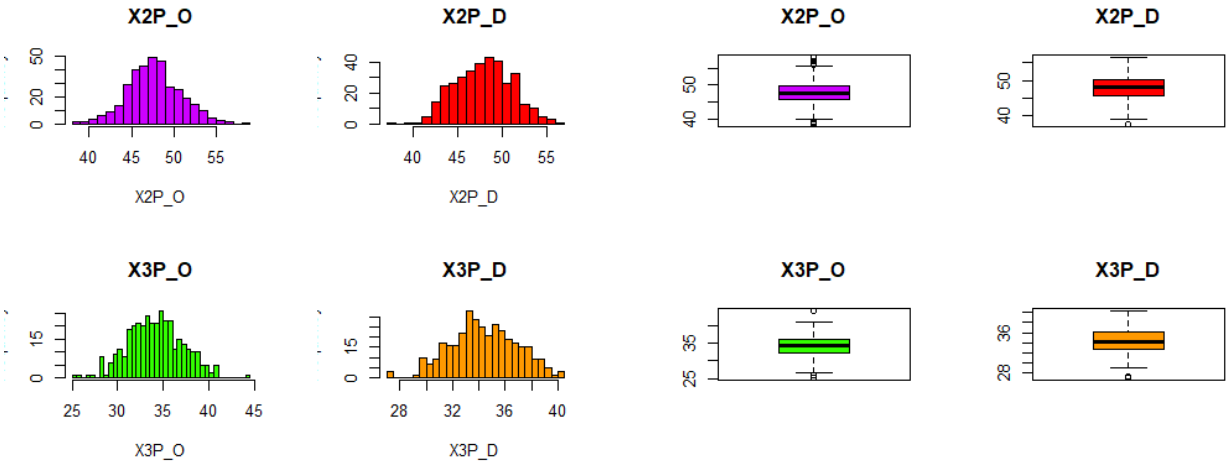
- These four features tell us about the so called "hustle stats": how much a team is physical and aggressive.



- These four features provide us with additional insights, such as "tempo", (which indicates how much a team prefers to rush a shot or take it slow), how much a team is good at free throws, and how much a team is lucky for the opponents not to score them.



- These four variables give us information about the percentages of two-point and three-point shots a team can score, and the percentage of shots a team concedes.



Also, another important thing we should know about the distributions of our features, is if they are skewed, so we report here the results (ordered descendingly):

Colonna	Skewness
TORD	0.44342100
ADJ_T	0.41833708
FTRD	0.39892387
W	0.38141197
G	0.37959171
WAB	0.33545458
ADJOE	0.26497906
FTR	0.25552995
TOR	0.24528947
X2P_O	0.14438602

Table 1.2: Skewness values - Part 1

Colonna	Skewness
BARTHAG	0.13968969
EFG_O	0.13234944
DRB	0.05709217
X3P_O	0.04602185
win_ratio	-0.01278230
X3P_D	-0.03491688
X2P_D	-0.03980490
EFG_D	-0.07761930
ADJDE	-0.10345452
ORB	-0.10587613

Table 1.3: Skewness values - Part 2

This will be important in the section about the feature selection for the PCA analysis.

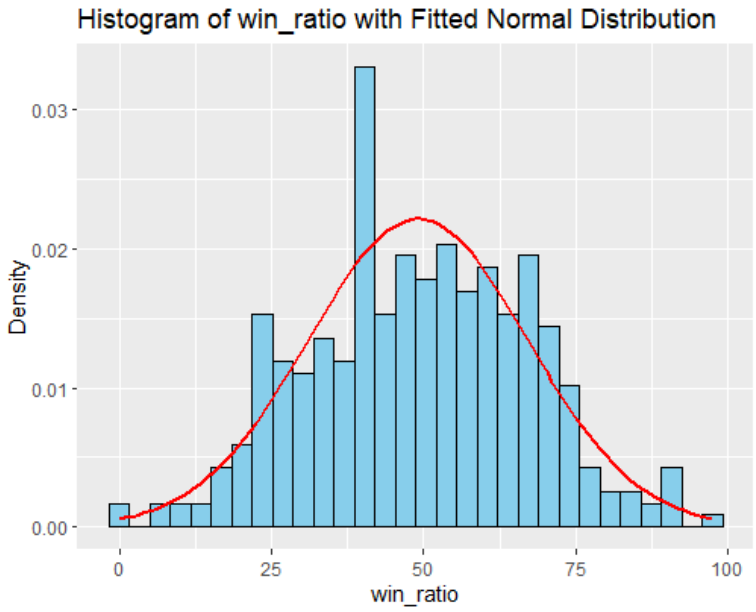
To conclude the EDA, we have a closer look at the win ratio, our newly own-made variable. It

presents a very low value of skewness, and from the histogram it appears to follow a normal distribution: if we perform a Shapiro-Wilk test, this produces a very high p-value (0.3634), so we cannot reject the null hypothesis (H_0 : the distribution is normal).

We then try and fit a normal distribution on it, getting the following result and plot:

Parametro	Valore
μ	49.0209189
σ^2	18.0306956

Table 1.4: μ e σ^2 for the win ratio fit



2 | Experimental Study

2.1. Regression Analysis

2.1.1. Feature Selection and Correlation Heatmap

One of the fundamental aspects of our analysis involves employing linear regression techniques to uncover relationships between various metrics and team success. Linear regression serves as a powerful tool for exploring the predictive potential of different factors on outcomes of interest.

In order to get an overview of possible features candidates for a linear dependence analysis we have a look at the correlation matrix through the correlation heatmap plot:

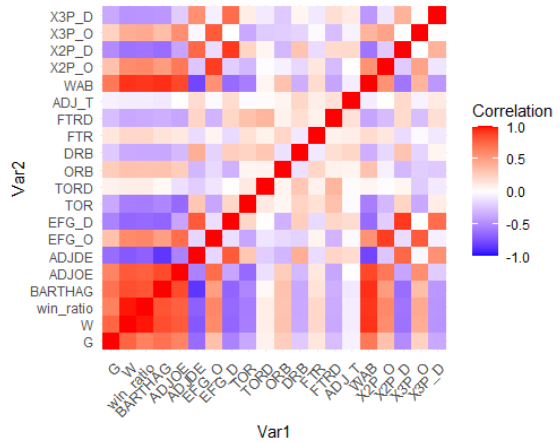


Figure 2.1: Heatmap of correlations

This visual aid offers valuable insights into the interaction between different variables, helping in the selection of suitable predictors for our regression model.

Obviously, we can see from the correlation heatmap that there are lots of highly correlated couples of variables, but only few of them seem to be interesting. In particular, we will be interested in analyzing our newly created variable "win ratio" to understand if there exist a significant linear relationship between some variables and a team's win percentage, encapsulating a team's overall performance throughout a season.

We can then have a look at the "win ratio" correlation row and also see the numerical values:

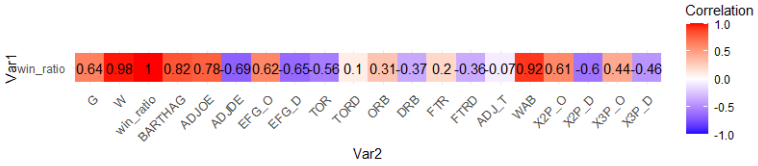


Figure 2.2: Win ratio correlations

We can see that, out of the non trivial ones, the feature which has the highest correlation with win ratio is ADJOE; thus, we will focus on understanding how offensive efficiency influences a team's ability to secure victories throughout a season.

At the same time, if we have a look at the scatterplot of such two features, we can clearly notice a strong linear relationship between the two selected features, in particular with a positive correlation index:

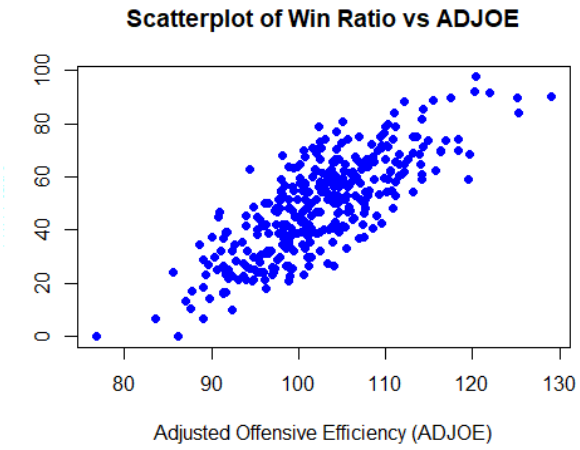


Figure 2.3: Scatterplot of the selected features win ratio and ADJOE

2.1.2. Least Squares and Residual Analysis

Ordinary Least Squares (OLS) is a fundamental technique used in regression analysis to estimate the relationship between a dependent variable and one or more independent variables. It works by minimizing the sum of the squared differences between the observed and predicted values of the dependent variable. The formula for Ordinary Least Squares is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

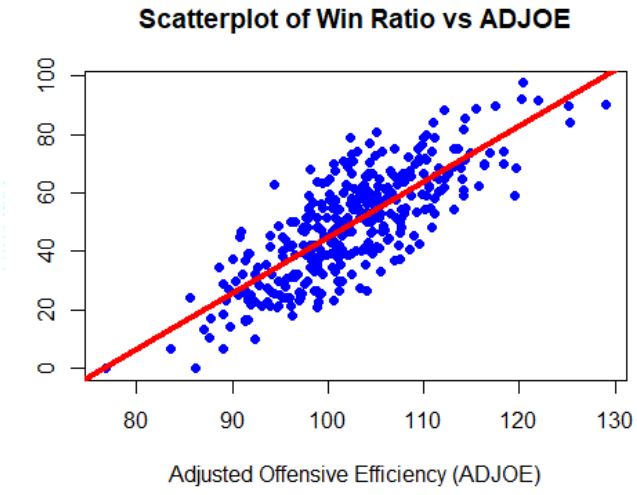
where:

- $\hat{\beta}$ represents the estimated coefficients
- X is the design matrix of independent variables
- Y is the vector of observed values of the dependent variable

We can then proceed to build a linear regression model to further explore the relationship between Adjusted Offensive Efficiency (ADJOE) and win ratio. Our model will take the form:

$$Y = \beta_0 + \beta_1 X$$

where X represents ADJOE, Y represents win ratio, β_0 is the intercept, and β_1 is the coefficient corresponding to ADJOE. Building this model we can easily find and plot the resulting regression line:



Let's also have a look at the normality plots for the standardized residuals:

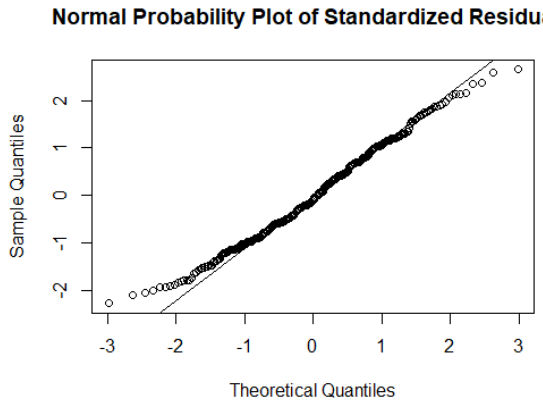


Figure 2.4: Normality qqplot of the standardized residuals from the linear model

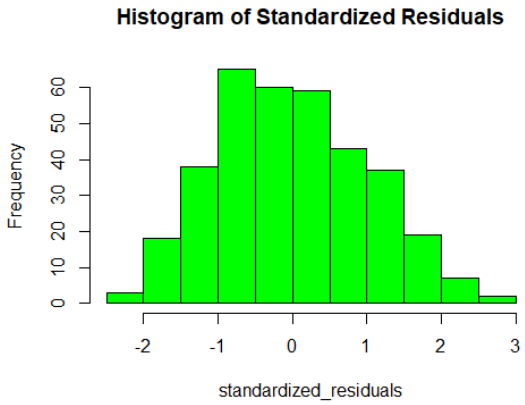


Figure 2.5: Histogram of standardized residuals

From looking at both of these plots, we see that the standardized residuals may appear slightly right-skewed; additionally, from the qqplot, we observe a minor "inverted S" pattern, indicative of a leptokurtic distribution. However, despite these observations, the assumption of normality can be deemed acceptable, as there is no strong evidence to reject it. Therefore, we can proceed with caution in interpreting the model without significantly compromising the reliability of the estimates obtained.

Lastly, for completeness, we report the value of skewness of the standardized residuals:

```
> skew <- skewness(residuals)
> print(skew)
[1] 0.1954581
```

2.1.3. Population Regression Equation

The population regression equation for the linear model is given by:

$$\text{win\_ratio} = \beta_0 + \beta_1 \times \text{ADJOE}$$

Interpreting the coefficients:

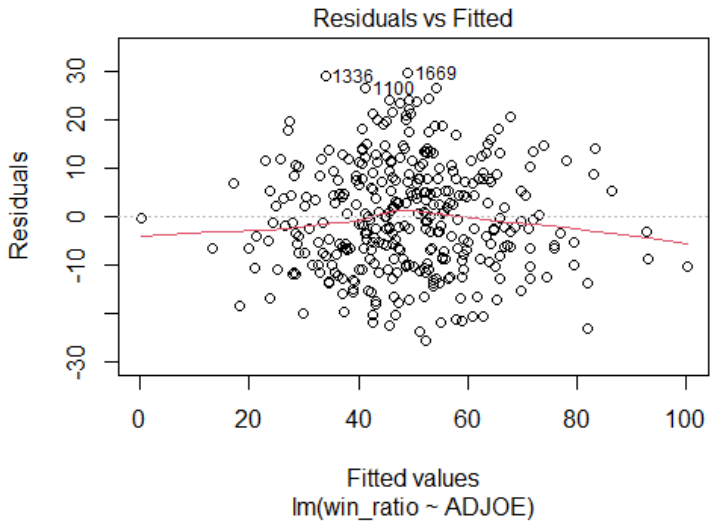
- β_0 (Intercept): This represents the estimated win ratio when the independent variable (ADJOE) is zero. In this case, it is estimated to be -146.1568,but it has not a very clear interpretation: there is no team that has a value ADJOE close to 0, and also it would not be possible to have a negative win ratio. We will later focus on this aspect
- β_1 (ADJOE): This represents the change in win ratio for a one-unit increase in the independent variable (ADJOE), holding all other variables constant. In this case, it is estimated to be 1.9076.

2.1.4. Verifying Regression Assumptions

In linear regression analysis, it is essential to verify the assumptions underlying the model to ensure the validity and reliability of the results. These assumptions provide the foundation for the interpretation of the regression coefficients and the predictive capabilities of the model. By confirming that these assumptions hold, we can have confidence in the accuracy of our regression analysis.

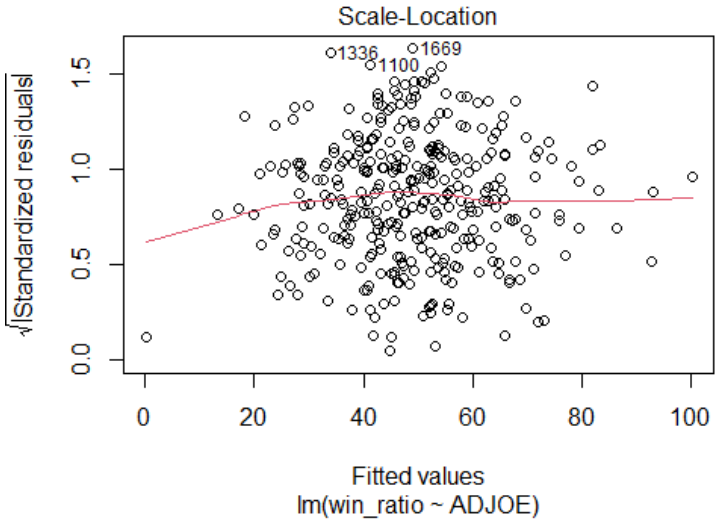
We already talked about the normality assumption of the residuals, so let's have a look at the other important hypothesis:

- **Zero mean assumption:** We need to confirm that the residuals have a mean of zero. Deviations from this assumption could indicate a bias in the model's predictions. To do this, we may look at the scatterplot of the residuals against the fitted values:



Since we can clearly see that there is no discernible pattern as the fitted values change, we confidently say that the zero mean assumption is respected.

- **Constant variance assumption:** We need to ensure that the variance of the residuals is constant across all levels of the predictor variables. Violations of this assumption may lead to inefficient parameter estimates. A very informative plot about this is the Square Root of Standardized Residuals vs Fitted Values plot (scale-location):



Also in this case, there in no visible pattern.

- **Independence assumption:** We need to verify that the residuals are independent of each other, meaning that the error terms are not correlated with each other. Autocorrelation in the residuals can lead to inefficient parameter estimates and invalid hypothesis tests. While there isn't a specific plot that directly confirms the independence of residuals, we can perform some statistical tests like the Durbin-Watson test and the Ljung-Box test for the independency of the residuals:

```
        Box-Ljung test

data:  residuals(lm_model_2015)
X-squared = 0.24077, df = 1, p-value = 0.6236
```

Figure 2.6: Ljung-Box test

```
 lag Autocorrelation D-W Statistic p-value
   1      0.02607943      1.945528   0.586
 Alternative hypothesis: rho != 0
```

Figure 2.7: Durbin-Watson test

Both the tests performs really well and both of them give a high p-value: we can then

conclude that we cannot reject the null hypothesis of autocorrelation for every significance level.

Now that we have verified that the assumptions of the linear regression model are met, we can have confidence in the coefficient estimates and the conclusions derived from the model. We can use the model to gain a better understanding of the relationship between the predictor variables and the response variable and to make predictions on new data or infer the effects of the predictor variables on the response variable

2.1.5. Correlation and Determinacy Coefficient

The correlation and determination coefficients are two important measures to consider when analyzing a linear model, let's delve into their specific meaning and interpretation:

- **Correlation coefficient (R):** The correlation coefficient measures the strength and direction of the linear relationship between the predictor variable (ADJOE) and the response variable (win_ratio). In this case, the correlation coefficient is approximately 0.782, indicating a strong positive linear relationship between the team's adjusted offensive efficiency (ADJOE) and their win ratio. This suggests that teams with higher adjusted offensive efficiency tend to have higher win ratios.
- **Coefficient of determination (R-squared):** The coefficient of determination represents the proportion of the variance in the response variable (win_ratio) that is explained by the predictor variable (ADJOE) in the linear model. In this case, the coefficient of determination is approximately 0.611, indicating that about 61.1% of the variability in the win ratio can be explained by the adjusted offensive efficiency (ADJOE) alone. This suggests that the linear model provides a reasonably good fit to the data, with ADJOE explaining a substantial portion of the variability in win ratio among teams.

2.1.6. Hypothesis Test of Linear Dependence

Statistical Hypothesis Test:

To determine whether a linear relationship exists between the chosen variables, we conduct a hypothesis test. The null hypothesis (H_0) states that there is no linear relationship, while the alternative hypothesis (H_1) states that there is a linear relationship.

Mathematically, the hypotheses can be stated as:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

In a linear regression model, testing whether a coefficient is equal to zero or not is typically done using a t-test. Specifically, if we are testing whether the coefficient β_1 is not equal to zero ($\beta_1 \neq 0$), we would conduct a two-tailed t-test. The test statistic for the t-test is calculated as:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ is the estimated coefficient, and $SE(\hat{\beta}_1)$ is the standard error of the coefficient estimate.

Under the null hypothesis ($H_0 : \beta_1 = 0$), the test statistic follows a t-distribution with $n - 2$ degrees of freedom, where n is the number of observations.

We can then calculate the p-value associated with the t-test to determine whether to reject or fail to reject the null hypothesis. If the p-value is less than your chosen significance level (e.g., 0.05), you will conclude that there is evidence of a significant linear relationship.

Results of the t-test for Coefficient β_1:

The t-test for the coefficient β_1 tests the hypothesis that there is no linear relationship between the predictor variable and the response variable. The results of the t-test are as follows:

- **Estimate**: The estimated value of the coefficient β_1 is 1.9076.
- **Standard Error**: The standard error associated with the estimate of β_1 is 0.0814.
- **t value**: The calculated t value for β_1 is 23.43.
- **p-value**: The p-value associated with the t-test for β_1 is less than 2×10^{-16}.

The highly significant p-value suggests strong evidence against the null hypothesis that the coefficient β_1 is equal to zero. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between the predictor variable and the response variable.

2.1.7. Confidence and Prediction Intervals

Construction and Interpretation of 95% Confidence Interval for Slope: To construct a 95% confidence interval for the unknown true slope of the regression line, we use the `confint()` function in R. The results of the confidence interval calculation are as follows:

	Lower Bound	Upper Bound
(Intercept)	-162.579931	-129.733702
ADJOE	1.747511	2.067706

Table 2.1: Intercept and ADJOE

The confidence interval for the slope of the regression line (corresponding to the variable `ADJOE`) is $[1.747511, 2.067706]$. This means that we are 95% confident that the true slope of the regression line falls within this interval.

Interpreting the confidence interval, we can conclude that there is a statistically significant linear relationship between the predictor variable (`ADJOE`) and the response variable at the 95% confidence level. Specifically, for each one-unit increase in `ADJOE`, we would expect the response variable to increase between 1.747511 and 2.067706 units, on average, while holding all other variables constant. This is, in our opinion, a strong results: adding 3 or 4 points to your ADJOE value results in an increase in wins of almost 8%, which can lead to a radical change from a good season to an excellent one.

This confidence interval provides valuable information about the precision of our estimate for the slope of the regression line and helps us make more informed conclusions about the relationship between the variables.

Construction and Interpretation of 95% Confidence Interval for Population Correlation Coefficient:

To construct a 95% confidence interval for the population correlation coefficient between the variables `ADJOE` and `win_ratio`, we use the `cor.test()` function in R. The results of the confidence interval calculation are as follows:

Confidence Interval: [0.7376465, 0.8195440]

This confidence interval means that we are 95% confident that the true population correlation coefficient between `ADJOE` and `win_ratio` falls within the range of 0.7376465 and 0.8195440.

Interpreting the confidence interval, we can conclude that there is a strong positive correlation between the variables `ADJOE` and `win_ratio` at the 95% confidence level.

Since the range of this interval is not very big, we are happy with the result of our analysis. The narrow width of the confidence interval indicates that our estimate of the population correlation coefficient is relatively precise. This gives us confidence in the strength and direction of the relationship between the variables `ADJOE` and `win_ratio`.

Based on this confidence interval, we can conclude that there is a statistically significant and strong positive correlation between the adjusted offensive efficiency (`ADJOE`) of basketball teams and their win ratios (`win_ratio`). This finding is consistent with our expectations and provides valuable insights into the factors that influence team success in basketball.

Overall, the confidence interval reinforces the validity of our analysis and supports the conclusions drawn from the data.

Construction and Interpretation of 95% Confidence Interval for the mean value of y, at fixed values of x:

To further understand the relationship between the predictor variable (`ADJOE`) and the response variable (`y`), we construct a 95% confidence interval for the mean of the response variable (`y`) at fixed values of `ADJOE`. This allows us to estimate the average value of `y` for a specific level of `ADJOE` with a certain level of confidence.

ADJOE Value	Fit	Lower Bound	Upper Bound
120	82.76	79.69	85.82
110	63.68	61.97	65.39
100	44.60	43.36	45.84
90	25.53	23.23	27.83
80	6.45	2.69	10.22

Table 2.2: Predicted Mean of y with 95% Confidence Interval

The table presents the predicted mean of the response variable (`y`) along with 95% confidence intervals for different values of the predictor variable (`ADJOE`). These intervals provide a range of values within which we can be 95% confident that the true mean of `y` lies for each corresponding

level of ADJOE.

Consider, for example, the ADJOE level of 110, which often characterizes a high-performing team. With 95% confidence, we can assert that the mean win ratio for teams at this ADJOE value lies within the range of 61.97 to 65.39. This interval provides valuable insight into the expected performance of teams with a pretty strong offensive efficiency.

Construction and Interpretation of 95% Prediction Interval for the value of y, at fixed values of x:

To evaluate the variability in individual predictions, we construct prediction intervals for the response variable (y) at fixed values of the predictor variable (X1). Prediction intervals provide a range of values within which we can expect future individual observations to fall with a certain level of confidence.

ADJOE Value	Fit	Lower Bound	Upper Bound
115	73.22	50.93	95.51
105	54.14	31.94	76.35
95	35.07	12.83	57.30
85	15.99	-6.38	38.36

Table 2.3: Predicted Value of y with 95% Prediction Interval

The table presents the predicted value of the response variable (y) along with 95% prediction intervals for different values of the predictor variable (X1). These intervals provide a range of values within which we can expect future individual observations of y to fall with 95% confidence.

Consider, for example, the ADJOE level of 115, which surely characterizes a high-performing team. With 95% confidence, we can assert that the win ratio for this team at this ADJOE value lies within the range of 50.93 to 95.51. This results is showing us how unpredictable sport can be: a team with 50.93% win ratio is just an average team, while a team with 95.91% win ratio is unforgettable.

Notably, that the prediction intervals (for a random value of y) are always bigger than the confidence intervals (for the mean value of y). We can also visualize this through two useful plots:

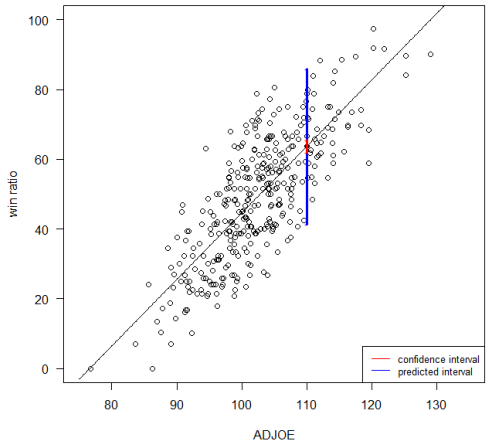


Figure 2.8: Intervals for fixed value of ADJOE=110

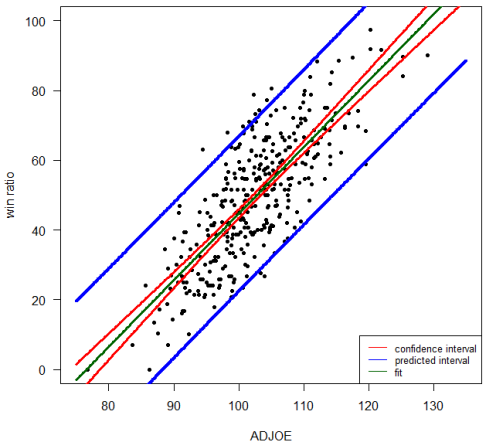
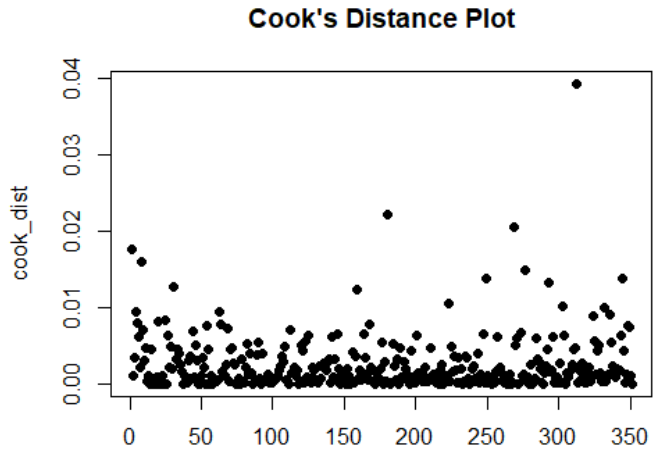


Figure 2.9: Pointwise intervals

2.1.8. Outliers and High Leverage Points

Outliers, high leverage points, and influential points are important considerations in linear regression analysis. These data points can significantly impact the results of the regression model, influencing parameter estimates, model fit, and the overall interpretation of the relationship between variables. Understanding and identifying these points is crucial for ensuring the validity and reliability of the regression analysis. Let's delve into each of these concepts to understand their implications in the context of linear regression.

Let's have a quick look at the influence of every point in the linear regression using the Cook distance:



As we can see from the plot, there is no point with a very high Cook's distance (every one of them is below 0.04).

> max(cook-distance) [1] 0.03935016

Anyway, we should take a look at the point with the maximum Cook's distance, since it almost has twice the value of the second largest one.

> which.max(cook-dist) 1569 312

This indicates us that is the point at row 1569 of the original dataset, indexed by 312 now (namely, Indiana University)

If we now have a look at the standardized residuals, we can also see that this one team also has one of the highest (absolute) values. This means that this could be a high influential point and an outlier.

However, if we have look at the leverage value for this point in the dataset, we have that it's very low; resuming:

- **Cook's distance:** 0.03935016, really below the indicative threshold of 1
- **Leverage:** 0.0182508, just a little more than the indicative theoretical threshold of $\frac{2p}{n}$.
- **Standardized residual:** -2.056, pretty low

This team was certainly unique: while demonstrating outstanding offensive efficiency, finishing the season ranked 25th in points scored per game and 9th out of 351 teams in the entire NCAA for three-point shooting percentage, Indiana University showed notable defensive vulnerabilities, even ranking 327th in points allowed per game. Thus, despite their offensive prowess, their defensive shortcomings frequently made them vulnerable to defeats, especially against stronger opponents.

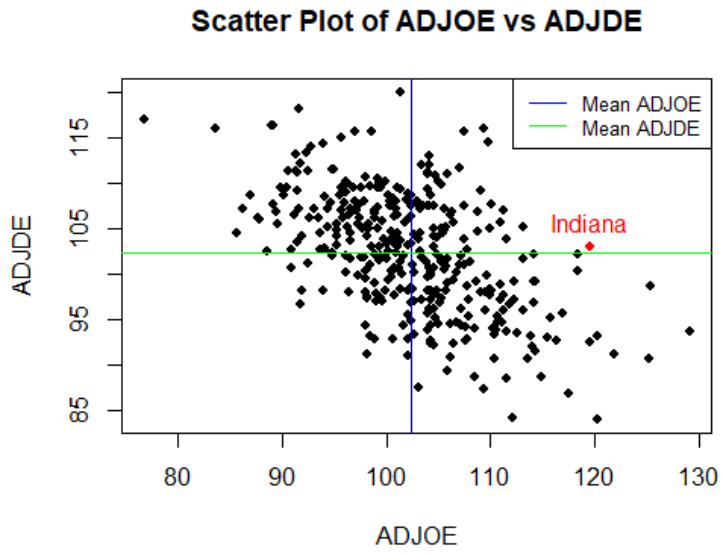
This team serves indeed as a clear illustration of a well-established concept in basketball: prioritizing scoring without sufficient emphasis on defense typically leads to limited success, especially on important games.

We can conclude that Indiana University may be considered an outlier in our model: the sole information we utilized (ADJOE) appears insufficient to explain this team's win ratio.

This observation also underscores an issue and a potential extension, which involves a linear model incorporating multiple variables as inputs. For instance, incorporating ADJDE would

enable us to consider defensive inputs alongside offensive ones and take care of team like Indiana University.

To visualize this, we can just have a look at the following plot:



2.2. Principal Component Analysis

2.2.1. Introduction

In our PCA, we have selected 6 specific features to delve into the underlying structure of team performance, namely Adjusted Offensive Efficiency (ADJOE), Adjusted Defensive Efficiency (ADJDE), Three-Point Shooting Percentage (X3PO), Two-Point Shooting Percentage (X2PO), Three-Point Shooting Percentage Allowed (X3PD) and Two-Point Shooting Percentage Allowed (X2PD).

Multiple factors guided the selection of these features. Firstly, our aim was to enhance interpretability, which led us to pair offensive and defensive components for each metric. This approach facilitates a more intuitive understanding of the results. Secondly, ADJOE was deemed essential due to its strong correlation with win percentages, as revealed by prior linear regression analysis.

Additionally, the inclusion of X2PO and X3PO was motivated by their ability to showcase a team's proficiency in orchestrating effective scoring plays. Lastly, the symmetry observed in these variables during exploratory data analysis underscores their suitability for PCA, ensuring clarity in interpretation.

By incorporating these features, we intended to uncover insights into the dynamics that shaped team performance in collegiate basketball during the 2015 season.

Let us have a look at the boxplot of the six features that we have chosen to perform Principal Component Analysis:

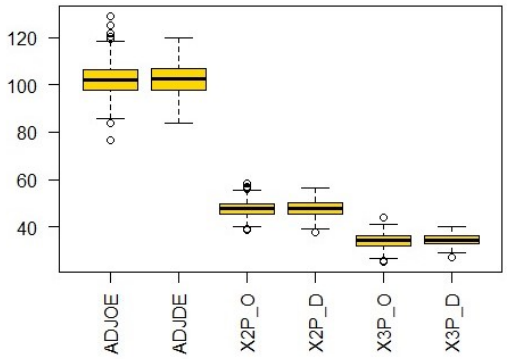


Figure 2.10: Boxplot

2.2.2. Explained Variance and Normalization

For our analysis, we performed Principal Component Analysis (PCA) over both the features normalized with z-score and by range.

Let us have a look at the two boxplots of our data over the six features we have chosen, normalized by z-score and by range:

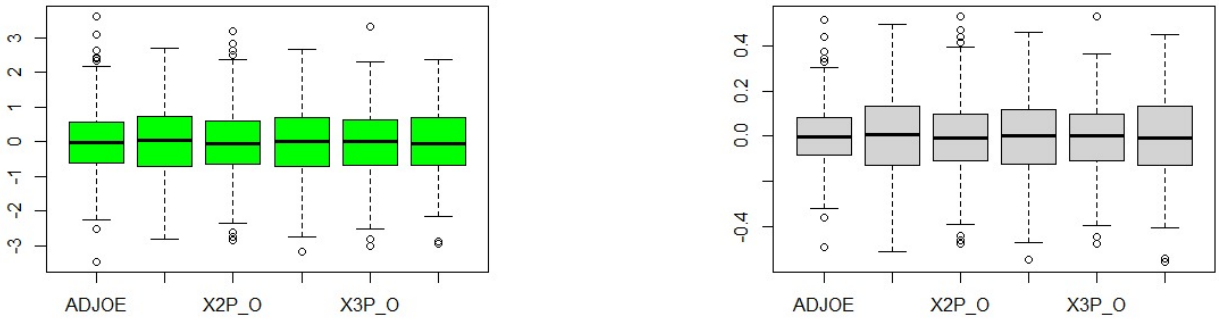


Figure 2.11: Boxplots of normalized data

As it appears in the two plots, it is reasonable to think that the normalization does not really make a difference in terms of the new features identified by the PCA.

Indeed, the cumulative explained variance confirms that the choice between the two normalization methods did not significantly impact the results, as both methods yield equivalent outcomes.

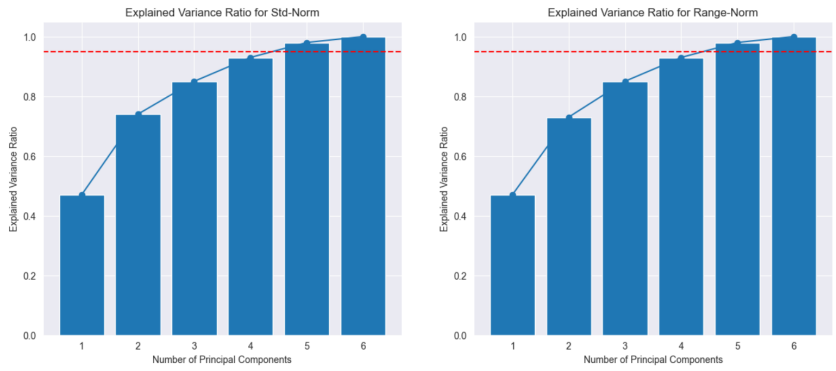


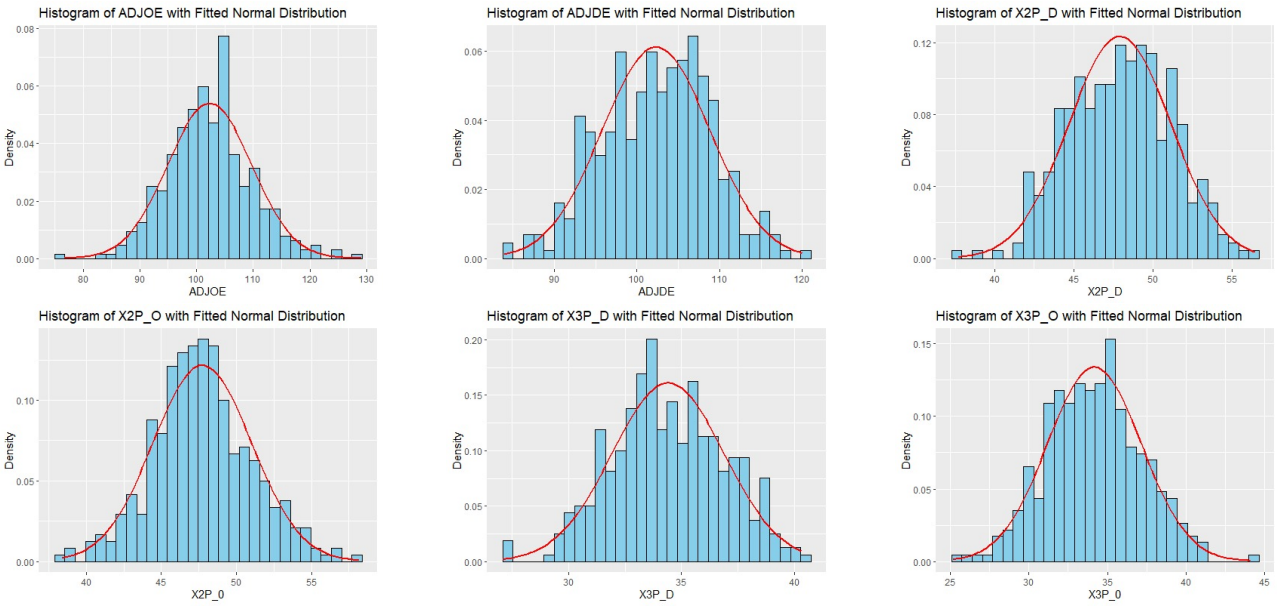
Figure 2.12: Explained variance for each normalization

However, we opted for z-score normalization over range normalization, guided by the preferable suitability of z-score normalization for features with unimodal distributions.

This choice was informed by conducting the Shapiro-Wilk normality test with a significance level of 5%, which aimed to assess the distributional characteristics of our selected features, and supported our hypothesis, indicating normality for these features, which implies unimodality.

To illustrate, we present distribution plots of the six selected features: ADJOE, ADJDE, X3PO, X2PO, X3PD, and X2PD.

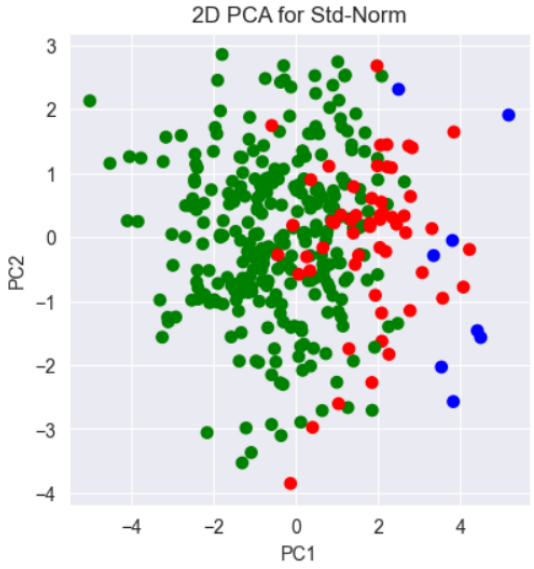
To illustrate, we present distribution plots of the six selected features:



2.2.3. Principal Components and Interpretation

Subsequently, we leveraged the information in the 'POSTSEASON' column to classify data points for visualization.

Assigning three distinct colors—representing "top teams," "average teams," and "bad teams" — facilitates the interpretation of performance tiers. Particularly, we plotted in blue the top 8 teams (out of 351, those who ranked top 8), in red the teams who got kicked out in Round of 64, Round of 32 and Sweet 16, and in green all the others who got kicked out earlier in the tournament.



The relationship between team rankings at the end of the season and the newly derived score concepts characterized by the principal components (PCs) is a compelling insight gleaned from our analysis.

It's notable that all features in the dataset were intuitively related to some aspect of team performance, with higher/lower values indicative of better/worse team performance. Therefore, it is not surprising that our PCA effectively identifies good and bad teams based on these performance metrics.

The classification of data points according to their end-of-season rankings, visualized through both 2D and 3D PCA plots, reveals a clear correspondence between team performance as captured by the PCs and their final standings, besides a few exceptions which will be discussed later, in paragraph 2.2.5. This alignment underscores the efficacy of our PCA in distilling the complex array of performance metrics into meaningful dimensions that reflect team success.

Thus, our analysis not only confirms the expected relationship between performance metrics and team rankings but also highlights the utility of PCA in extracting actionable insights from multifaceted datasets in collegiate basketball analysis.

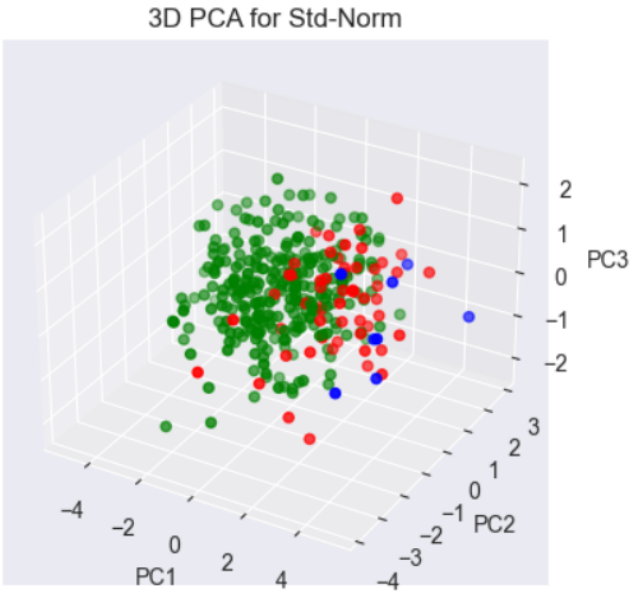


Figure 2.13: 3D PCA for Std-Norm

Lastly, in the next figure the six eigenvectors are depicted, to delve into the composition of the principal components and deeply understand the underlying structure of the data and the contribution of each feature to these components.

These plots offer straightforward interpretation:

- The first principal component represents a weighted sum of the six features, with positive weights assigned to offensive aspects and negative weights to defensive ones. Consequently, teams excelling in both offense and defense garner higher values in this component.
- Conversely to the first one, the second principal component features negative weights across all features, favoring teams proficient in Three-Point Shooting. This principal component is about preferring offense over defense, or viceversa.
- The third principal component highlights teams which manage to prevent Two-Point Shootings more than Three-Point Shootings, as it primarily yields positive values; those team with a high value of the fourth principal component will then be the ones that are less physical in the area and concede more outside shooting to the opposite teams.
- The fourth component favors teams proficient in Three-Point Shootings over Two-Point Shootings. The basketball revolution has influenced lot of teams in the last years to more

from three than from inside the area, so we can say that, in a certain way, teams with a high value of the fourth principal component will be the most modern ones.

This interpretation of the loadings of the principal components is a strong result. The playing style is something that every coach tries to instill in his team and which influences a lot the success of the organization, and PCA helped us discover interesting insights about it.

Overall, these 4 first principal components explain 93% of the variance of the data, as it was shown shown before.

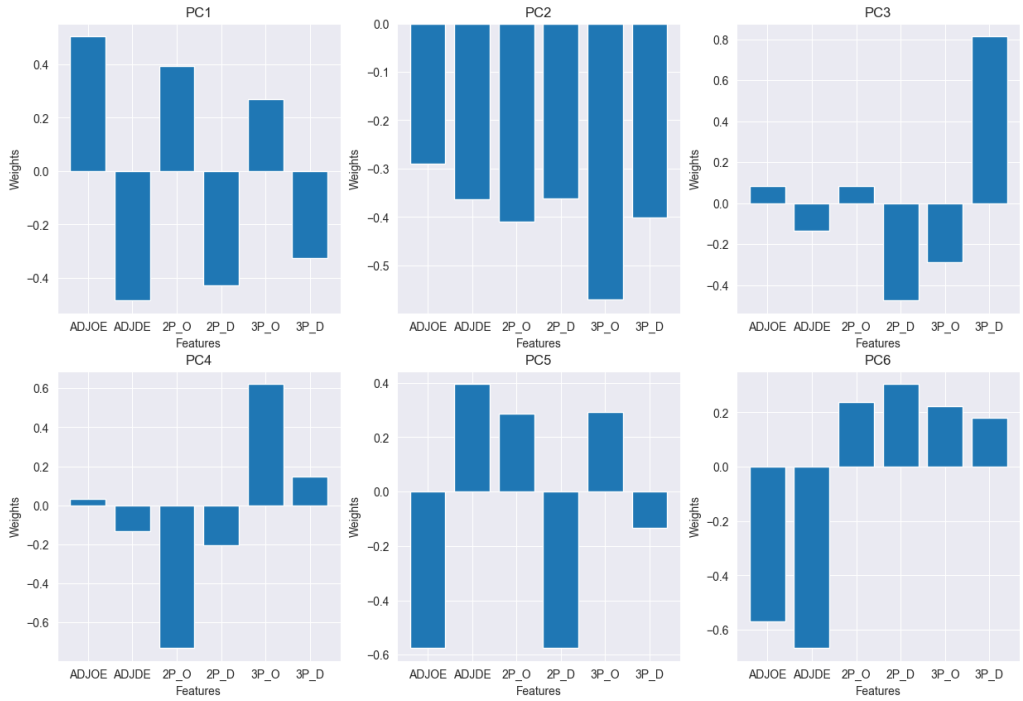


Figure 2.14: PCs Loadings

2.2.4. Singular Value Decomposition

Although Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) generally yield similar results, we conducted both methods to gain further insights into the underlying score concepts.

Upon decomposing the matrix into its constituent components — U, S, and $V^\intercal$ — we initially observed the rows of $V^\intercal$ and confirmed that SVD and PCA led to identical new features.

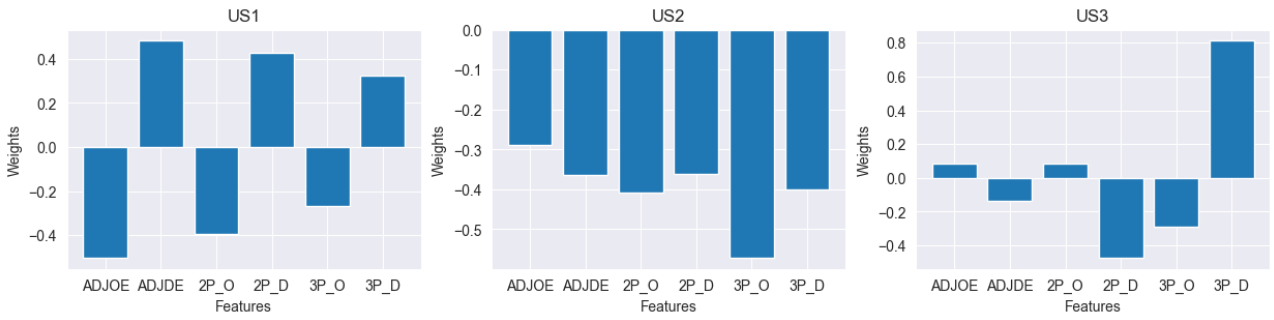


Figure 2.15: SVD Loadings

Subsequently, examining the matrix S, we discerned that the first new feature encapsulates a distinct concept characterized by significantly higher magnitude compared to the others.

```
# Print first two columns of u, s and v
s[:]
```
[28] ✓ 0.0s

```
array([489.77310826,  10.34819406,   7.04563958,   5.63547555,
         3.81840624,   3.14284165])
```

Figure 2.16: Concepts Strength

Thus, we decided to further investigate this one new feature.

Upon filtering the dataset and scrutinizing the values of the top 8 teams (depicted as blue points in the plots), we observed a consistent trend: these values were markedly negative and displayed remarkable uniformity.

```
# Print first column of U
u[:, 0]
```
✓ 0.0s

```
array([-0.37166022, -0.36306612, -0.32698824, -0.37439473, -0.34427705,
       -0.35829415, -0.34799157, -0.33905668])
```

Figure 2.17: Negative Talent

Therefore, analysing the SVD of our dataset, we got to the conclusion that the first new feature indeed encapsulates a concept of "negative talent," which is notably low among the top-performing teams. Notably, in the previous paragraph about the PCA, such concept was a "positive talent" score, but it was still equivalent.

2.2.5. A weird phenomenon

The phenomenon of upsets in sports, particularly in tournaments like the NCAA basketball championship, often defies expectations and challenges our preconceived notions of team performance.

As it was anticipated in paragraph 2.2.3, there are a few teams, like Hampton, UAB, Texas Southern, Harvard, San Diego St., North Dakota St., West Virginia, and Robert Morris, that seem to have outperformed expectations based on their data-driven classification.

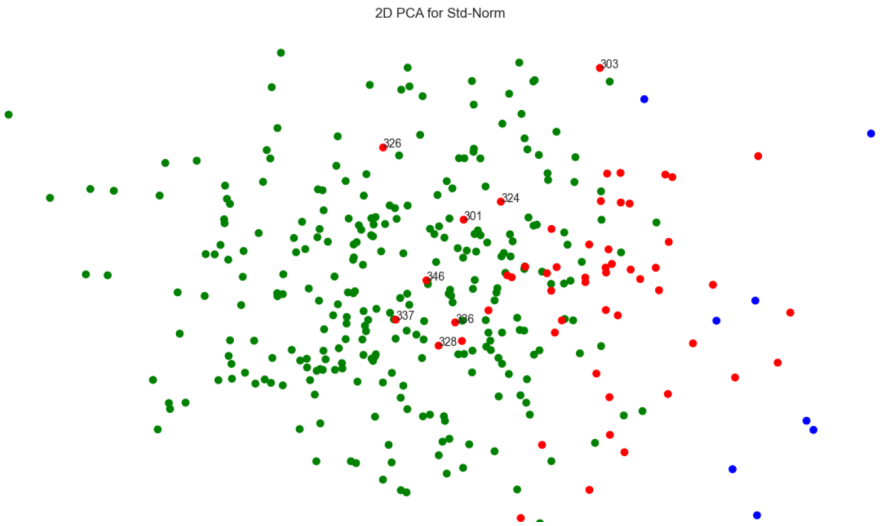


Figure 2.18: Outperforming teams

These teams, though initially classified within the cluster of perceived weaker teams, have surpassed expectations by achieving notable successes in the tournament. Such upsets occur when underdog teams, often with lower rankings or perceived as less competitive based on regular-season performance metrics, manage to defeat higher-seeded or more favored opponents.

In the case of the NCAA tournament, upsets inject excitement and unpredictability into the competition, capturing the essence of "March Madness". For instance, some of these teams may have advanced further in the tournament than anticipated, defeating higher-seeded opponents and garnering attention for their unexpected achievements.

We decided to have a close look at one of these upsets, one of the most unpredictable: on the 9th of march 2015, UAB, a really low ranked team, managed to get a win against the higher rated Iowa State University. In figure 2.18, UAB is depicted with index 301, evidently appearing in the "bad" teams class while being plotted as a red point, meaning that it managed

to achieve unexpected results. This event stands as a testament to the unpredictable nature of sports, a modern-day personification of the timeless tale of David versus Goliath.

More can be found... at this link.

This underscores the unpredictable and thrilling nature of sports, where statistical analysis and data-driven predictions can sometimes fall short in capturing the full spectrum of team capabilities and potential tournament outcomes.

2.3. Fuzzy Clustering

2.3.1. Introduction

In pursuit of a deeper comprehension of data features within the clustering process, we turned to the methodology of Fuzzy Clustering. Central to our investigation was the clustering of teams based on their performance metrics.

Indeed, with metrics ranging from offensive and defensive efficiency to shooting percentages and turnover rates, the dataset encapsulates the multifaceted dynamics that shape team success on the court. Through advanced analytical techniques such as fuzzy clustering, we have the opportunity to uncover nuanced insights into the diverse strategies, playing styles, and performance profiles of NCAA basketball teams. By leveraging fuzzy clustering's ability to capture complex relationships and account for ambiguity in cluster assignment, we can gain a deeper understanding of the underlying patterns and trends driving success in collegiate basketball during this period.

Upon visualizing the data and employing distinct color assignments to signify clusters, we discerned a conspicuous divergence among teams based on these pivotal features, successfully delineating distinct performance profiles among the teams under examination.

2.3.2. Fuzzy C-Means Clustering

For visualization, our attention was at first drawn to two key metrics, namely Adjusted Offensive Efficiency (ADJOE) and Adjusted Defensive Efficiency (ADJDE), which we had previously identified as robust indicators of team performance in chapter 2.1, where we delved into linear regression analysis.

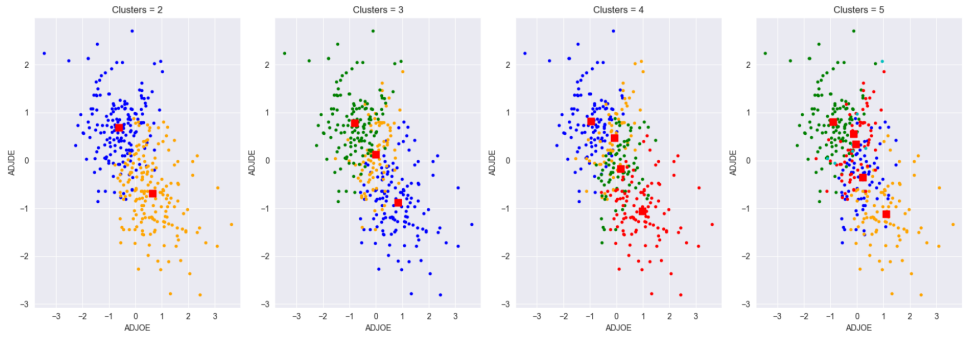


Figure 2.19: Fuzzy clusters over ADJOE and ADJDE

This segregation in the clusters corresponded to our initial hypotheses, unveiling distinct groupings that could be interpreted as indicative of both 'good' and 'bad' teams across a varied spectrum of values for c, ranging from 2 to 5.

Following this, we extended our analysis by projecting these identified clusters onto the first two principal components.

As elaborated upon during our exploration of principal component analysis, these first two principal components encapsulate significant variations in the data and offer meaningful insights into the underlying structure of the dataset.

Indeed, upon consideration of the underlying meaning attributed to the first two principal components, as previously identified in a comprehensive analysis outlined in section 2.2, it becomes evident that these clusters genuinely reflect the performance dynamics of the teams under scrutiny.

This observation further reinforces the validity and relevance of the identified clusters in characterizing the performance spectrum exhibited by the teams.

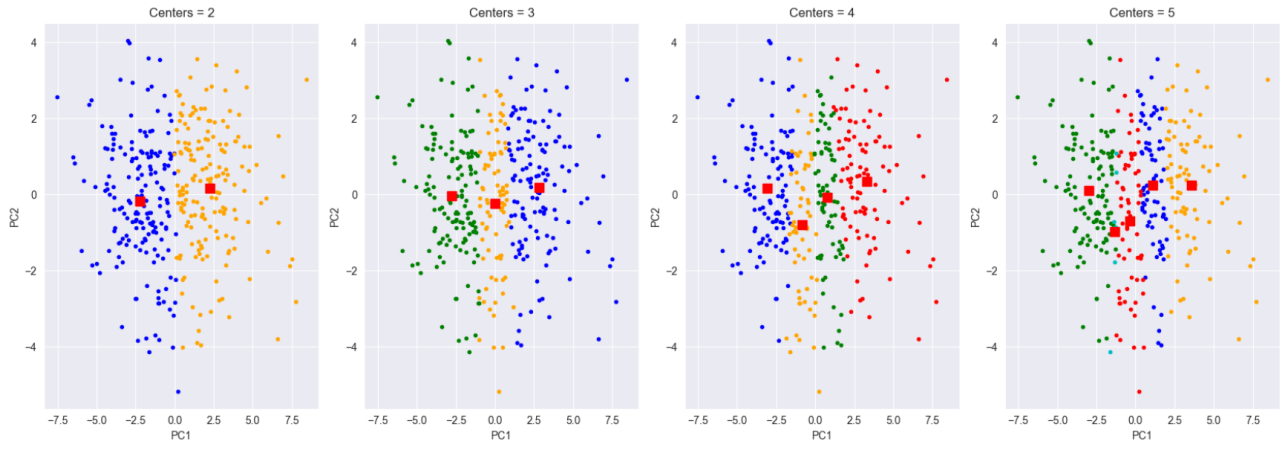


Figure 2.20: Fuzzy clusters over first two PCs

In our quest to determine the optimal number of clusters, we conducted an examination of the minimized cost function value for each value of c.

It is noteworthy that the cost function demonstrates a consistent trend of decrease with an increase in the number of clusters, as vividly depicted in the accompanying image below:

Figure 2.21: Descending cost function

Therefore, we opted for a more comprehensive approach by executing the Fuzzy C-Means Clustering algorithm with diverse random initialization centroids for 20 iterations, ensuring thorough exploration of the clustering landscape.

Subsequently, we assessed the stability of clustering quality across multiple iterations for each value of c.

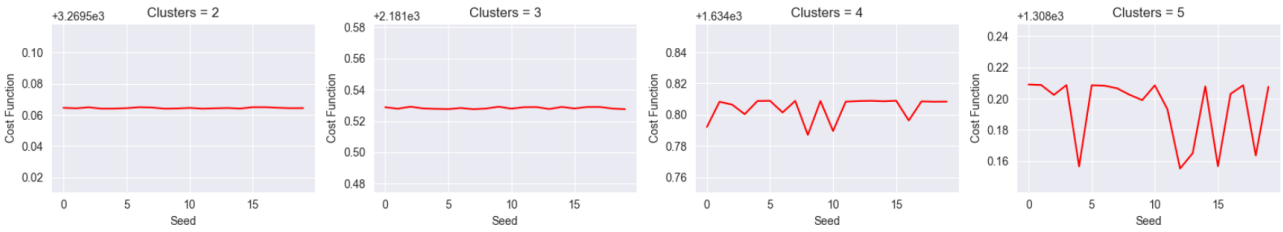


Figure 2.22: Stability of cost function

Upon observation, we noted that the clustering quality exhibited notably higher stability and resilience across iterations for configurations with 2 and 3 clusters. It suggests a greater degree of independence from the random initialization process, indicating a propensity towards convergence towards potentially optimal solutions.

The consistency in clustering outcomes for configurations with 2 and 3 clusters underscores their robustness and reliability in capturing the underlying structure of the data. Consequently, these configurations hold a starting clue in yielding the most meaningful and interpretable clustering outcomes for our analytical objectives.

However, before making further definitive conclusions, it is essential to highlight a crucial insight we gleaned from our examination, concerning the degree of membership of data points

with respect to all the clusters.

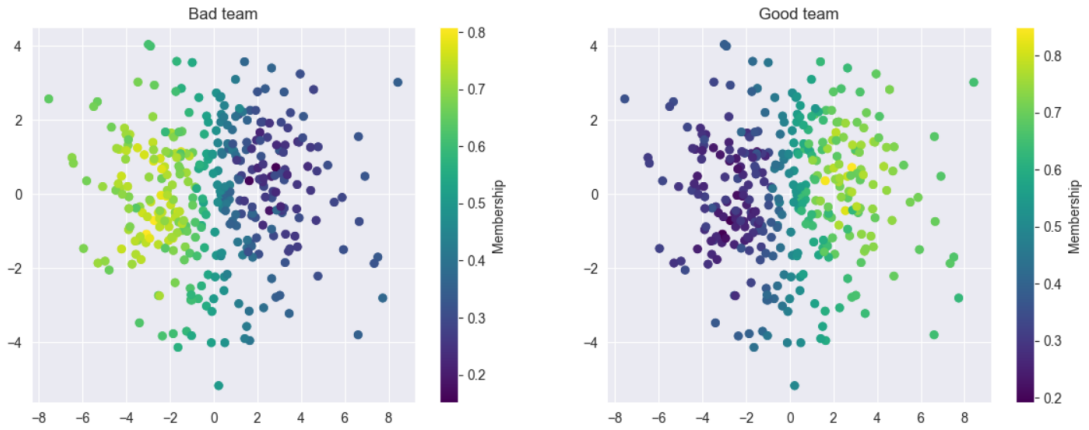


Figure 2.23: Membership plot with two clusters

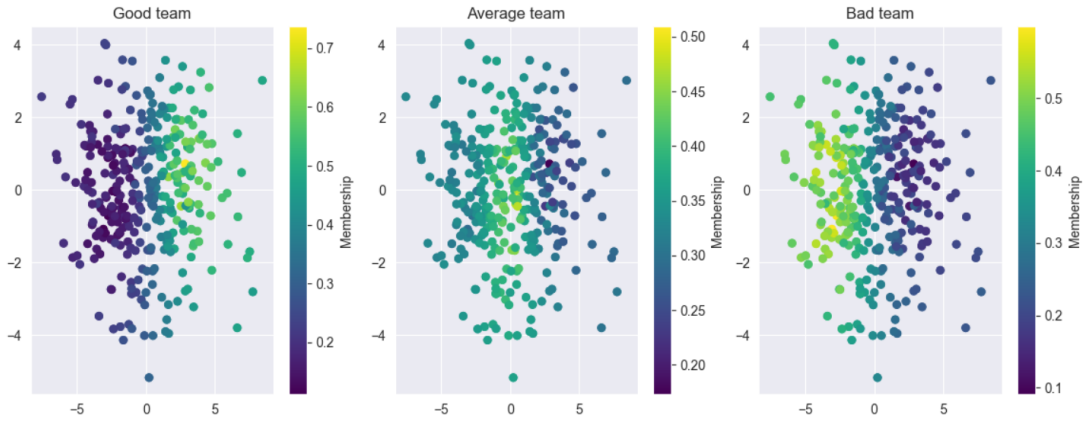


Figure 2.24: Membership plot with three clusters

In the examination of 2 clusters, a distinct and symmetrical partition emerged, delineating what could be construed as 'good' and 'bad' teams with evident clarity.

Meanwhile, upon expanding our analysis to encompass 3 clusters, the symmetry among clusters has been lost.

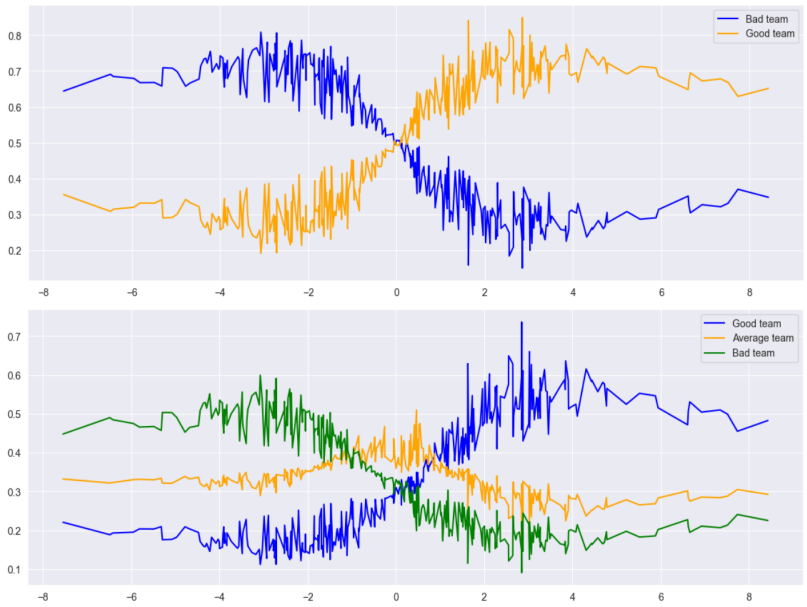


Figure 2.25: Membership values with two and three clusters

Notably, 'good' teams demonstrated a robust affiliation with their respective cluster, characterized by a cohesive grouping of performance metrics indicative of excellence. Conversely, 'bad' teams exhibited a more dispersed distribution, displaying a higher degree of association with the 'average' team cluster. This observation highlights the inherent challenge in precisely defining and categorizing 'bad' teams, underscoring the inherently ambiguous nature of such classifications within the clustering framework.

This nuanced exploration, as will be elaborated upon, once again highlights the crucial importance of our earlier deliberation on the selection of the parameter c.

As it was anticipated, our analysis will now delve deeper into exploring additional criteria and considerations to inform our selection of the most appropriate value for c, aiming to refine our understanding and arrive at a judicious decision regarding the optimal configuration for our clustering objectives.

2.3.3. Exploiting Anomalous Pattern Clustering

The Anomalous Pattern Clustering (AP-Clustering) method serves as an initial strategy to pinpoint the optimal value for c in our clustering analysis. It offers a systematic approach by detecting anomalous patterns within the dataset, a technique further utilized in AP-FCM

(Anomalous Pattern Fuzzy C-Means).

The first step we conducted in order to perform AP clustering was to execute it across a wide range of threshold values, aiming at analysing diverse configurations.

Given our dataset includes information on 351 entities, we opted to analyze all potential threshold options within 10 and 50. This approach ensured we would consider all values that might otherwise be overlooked, as depicted in the accompanying image below:

```
IAP identified 6 clusters with threshold of 10 elements
IAP identified 5 clusters with threshold of 15 elements
IAP identified 4 clusters with threshold of 22 elements
IAP identified 3 clusters with threshold of 33 elements
IAP identified 2 clusters with threshold of 38 elements
Reminder: the number of elements is  351
```

Figure 2.26: AP Threshold

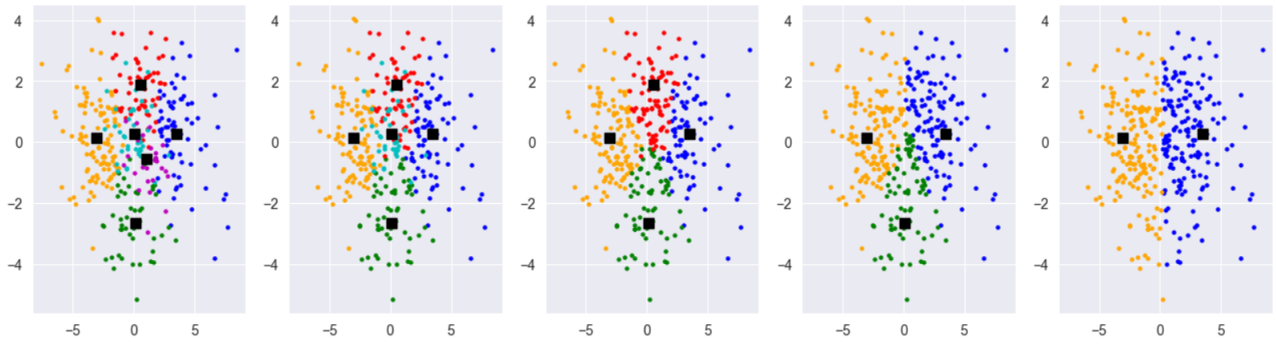


Figure 2.27: AP Partitions

With a quick stare at the image above, it is possible to tell that the most meaningful partitions happen thresholds from at least 15 data points.

Given this, we employed the partitions obtained for each threshold value to initialize the Fuzzy C-Means Clustering algorithm.

Remarkably, our findings revealed that the clustered data obtained in each case exhibited a striking resemblance to the outcomes obtained earlier.

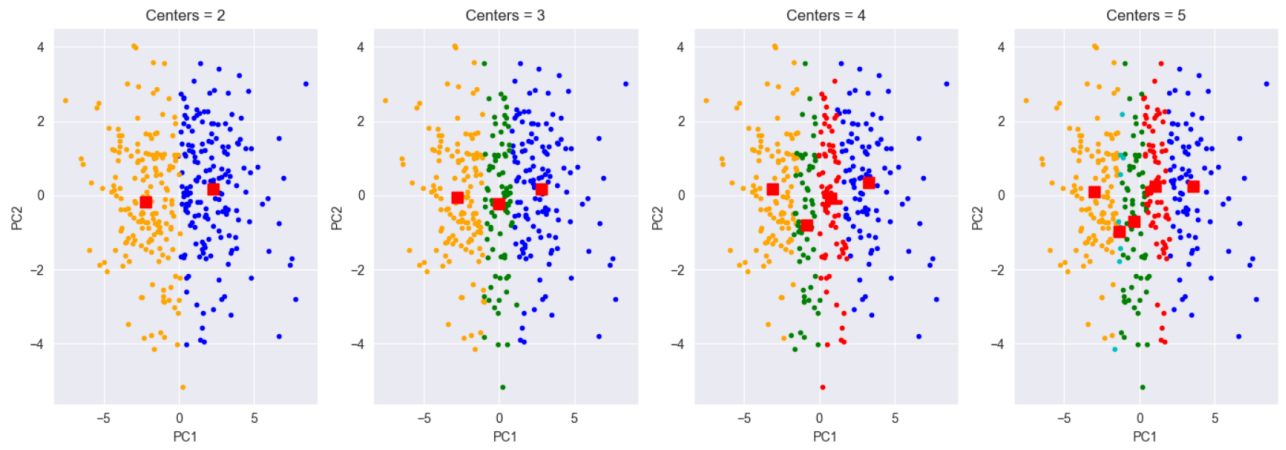


Figure 2.28: AP-FCM Clustering

Afterward, we proceeded to compare the minimum values of the cost function obtained from the Iterative Anomalous Pattern (IAP) clustering approach with those obtained through random initialization repeated over 20 iterations.

The results of this comparison, as depicted in the figure below, serve to reaffirm our earlier observations.

Specifically, they highlight once again that the cost function is most likely to be minimized towards a global minimum when the parameter c equals 2 and 3.

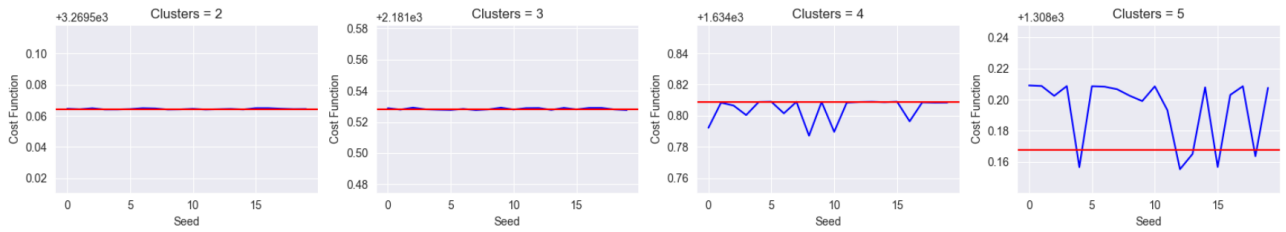


Figure 2.29: IAP Cost

The consistency observed across methodologies underscores the robustness and reliability of our findings, further validating the selection of these specific values of c in our clustering analysis.

To summarize, the outcomes derived from employing the Anomalous Pattern Clustering (AP-Clustering) technique to initialize the Fuzzy C-Means Clustering algorithm closely align with our earlier findings. This alignment instills greater confidence in the credibility of our results, a sentiment that will be reaffirmed in the subsequent paragraph.

2.3.4. Quality of clusters

Our quest for determining the optimal value for c continues indeed, as we delve into additional criteria and considerations. Particularly, we are leveraging the indexes implemented in the Fuzzy Clustering GitHub repository [6] to bolster our analysis.

Through rigorous experimentation, we executed the code for both the partitions obtained with traditional Fuzzy C-Means (FCM) and those initialized with Anomalous Pattern Fuzzy C-Means (AP-FCM).

Remarkably, our efforts yielded consistent results across both methodologies, further substantiating the reliability and robustness of our findings.

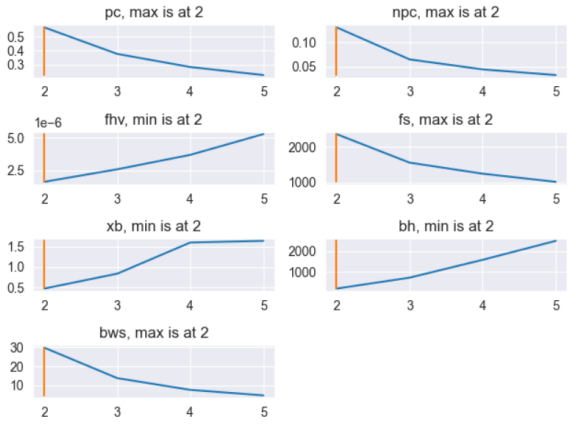


Figure 2.30: Indexes FCM

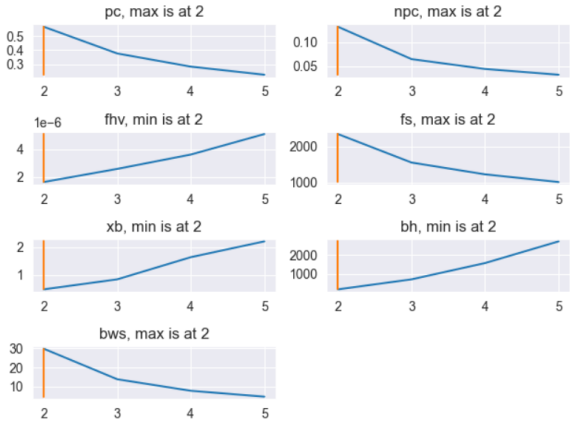


Figure 2.31: Indexes AP-FCM

At this juncture, it becomes apparent that the most meaningful value for c in Fuzzy C-Means Clustering is 2. Through various criteria and methodologies, including Anomalous Pattern Clustering and diverse evaluation metrics, we consistently converge on the conclusion that c=2 optimally captures the underlying structure of the data.

2.3.5. Interpretation

As a primary conclusion, it is evident that clusters effectively differentiate between teams of varying performance levels, with features serving as robust indicators. Notably, there is a clear ordering of clusters along the first principal component, which aligns with a gradient from negative to positive values, suggesting a continuum from "bad" to "good" teams for any value of c in Fuzzy C-Means Clustering.

Despite the desire within the sports science community for a three-cluster partition representing "bad," "average," and "good" teams, the analysis consistently indicates an optimal number of clusters at two. This finding underscores the challenge in defining what constitutes a "bad" team, as observed in the membership plots. Indeed, when considering the complexities of competitive dynamics, while it is straightforward to categorize teams with strong performance records as "good," defining "bad" teams proves elusive, especially considering factors such as early eliminations.

Ultimately, it transpires that the most significant partitioning of teams emerges along the axis of the first principal component, wherein teams are centered on their aggregated features, reflecting a more-or-less symmetrical division into two categories, with the center at value 0.

3 | Conclusion

In this chapter, we summarize the main findings of our analysis and discuss the challenges encountered during the research process. We also propose potential avenues for future investigation.

3.1. Main Findings

During our analysis, we encountered several challenges that may justify further investigation:

1. The normality of residuals for the linear regression model was found to be sufficient, but not particularly good. Despite attempting log-log transformation as requested, as well as Box-Cox transformation, no significant improvement was observed.

2. The linear model developed showed promising results, but certain anomalies were noted, as explained in paragraph 2.1.9. Specifically, teams like Indiana University exhibited very high offensive efficiency (ADJOE) but poor defensive performance, which deviated significantly from the overall trend. The presence of this outlier suggests that a linear model with additional variables could potentially provide more accurate predictions.

3.2. Future Directions

Based on our findings, we propose to conduct statistical analysis on time series data, in order to compare results across different years. By examining iterated trends, we may gain insights into long-term patterns and dynamics within college basketball.

In conclusion, while our analysis has illuminated several aspects of college basketball analytics, there remain areas that still require further exploration.

Indeed, we believe that, addressing the identified challenges, we can continue to advance our

understanding of the game and improve predictive models.

4 | Appendix

4.1. Scripts Index

This appendix provides additional details and resources related to the methodologies and tools used in the project. It includes scripts and notebooks used for data analysis and modeling:

- Scripts/Python/linear_regression.ipynb:
  Python notebook containing the implementation of linear regression
- Scripts/Python/pca.ipynb:
  Python notebook demonstrating Principal Component Analysis (PCA)
- Scripts/Python/scikit_fuzzy_clustering.ipynb:
  Python notebook showcasing fuzzy clustering using scikit-fuzzy
- Scripts/R/linear_regression.R:
  R script for performing linear regression analysis
- Scripts/R/pca.R:
  R script for conducting Principal Component Analysis (PCA)
- Scripts/R/eda.R:
  R script for Exploratory Data Analysis (EDA)

These scripts and notebooks serve as supplementary materials to the main project documentation, providing detailed insights into the data analysis and modeling techniques employed.

5 | Bibliography

5.1. References

[1] Vasant Dhar (2013) Data Science and Prediction Communications of the ACM.

[2] L'Heureux et al. (2017) Machine Learning With Big Data: Challenges and Approaches.

[3] Qiu et al. (2016), A Survey of Machine Learning for Big Data Processing, EURASIP Journal on Advances in Signal Processing 2016.

[4] A. Fahad et al. (2014) A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis, IEEE.

[5] M. Chen, S. Mao, and Y. Liu (2014): Big Data: A Survey.

[6] GitHub Repo https://github.com/99991/FuzzyClustering