Data and Basketball: An analysis of NBA Players' performance in the 2016-2017 season

Technical Paper

Adam Magyar, LaShawn Murray, Xiaotong He, Tianyi Tan and Yue You

Group liaison: Xiaotong He

March 22, 2019

Abstract

The NBA Player Social Power dataset has 446 cases, representing players in the league and social power is measured across 38 variables. Five analytical techniques were used to analyze these characteristics of player performance. These techniques included Principal Component Analysis (PCA), Common Factor Analysis (CFA), Canonical Correlation Analysis (CCA), Clustering Analysis and Ridge Regression. Common threads reflective of offensive and defensive performance indicators were highlighted across multiple analyses. This suggests that these indicators reflect important underlying performance statistics for NBA players. Further analysis is required to extend the analyses and find further insights as to player performance.

Introduction

The NBA Player Social Power dataset has 446 cases, representing players in the league and social power is measured across 37 variables. These variables are described in **Appendix B**. The data was sourced from kaggle.com at https://www.kaggle.com/noahgift/social-power-nba. In addition to offensive and defensive statistics, the dataset also contained information on parameters of interest include *the number of games played where the team won (W), rank (RK), and player impact factor (PIE)*. Salary information was integrated with the NBA Player Social Power dataset resulting in 39 variables across the 446 players.

Initial Data Analysis

The NBA dataset was comprised of both categorical, ordinal and metric variables. Variables of interest included the individual player, their position, salary, player impact factor and variables relating to player offensive and defensive performance (i.e. rebounds, fouls, and field goals). The ordinal variable, player rank (Rk) was of interest as it would highlight how the impact of the analytical techniques on the NBA's top players. In addition to on court observed statistics, the dataset also included statistics on players' relative skillset versus other players. These include offensive rank, defensive rank, wins attributable to a player, and a measure of how much better or worse a player is compared to the "average" player. This relative player statistic was called Player Impact Estimate (PIE). Those were all advanced NBA statistics provided by ESPN.

Several of these statistics could be redundant. For example, the dataset had a Total Field Goals, Total Field Goals Attempted, and Field Goal Percentage. The dataset also had 2 Point Field Goals, 3 Point Field Goals, 2 Pointers Attempted, 3 Pointers Attempted, 2 Point Percentage and 3 Point Percentage. Some the analysis excluded Total Rebounds and included Offensive and Defensive Rebounds to reserve

more information. Points were excluded from PCA and CFA for its calculation was directly related to other variables.

Missing Values

This dataset integrated the salary of each player with their season statistics. Consequently, the resulting dataset had 111 missing values for salary, as well as 33 missing values for 3P% (three-point field goal percentage) and 8 missing values for FT% (free throw percentage). When further evaluating these variables, it was assumed that they were blank because these players didn't attempt three pointers or didn't attempt free throws. To illustrate, one pattern found was that many of these blanks related to the position of Center (C), which based on this role, does not necessarily contribute to their three-point attempts. For free throw percentage, missing data was related to having very few minutes played. So we replaced the N/A in 3P% and FT% with zeros, based on the calculation. To resolve the missing values for salary, we consulted additional sources such as Hoopshype to find the player's salary.

Research Questions

Research Questions	**Methods**
To reduce the dimensionality and discover latent variables that allow players to be grouped by the style of how they play the game.	PCA
To reveal 2 factors (offensive and defensive) consistent with past research and our understanding on the nature of the factors. Can NBA players be segmented into groups based on their similarities in latent factor scores? What's the difference in average salary, wins and rank in each cluster?	SEM, CFA, Clustering Analysis using CFA's scores
To discover the association between player's skills set and player's performance and the difference between players with different salary levels.	CCA
How can NBA players be segmented into groups based on their similarities in indicators of performance?	Clustering Analysis
To shrink coefficients to non-zero value to reduce model complexity and standard errors. And to discover the variables (actions and performance) that impact the wins and points each player that result in the rank.	Multiple Linear Regression & Ridge Regression

Literature Review

Principal component analysis has been performed on sports studies. A research of Yin (2014) assigned an ability score to players based on several different statistics. PCA was done with an orthogonal rotation. Three principal components were found that could be used to define distinct attributes that a player may have. The first reflecting on scoring, assists, steals and faults. The author called this the "attack" factor as these were seen as aggressive moves by a player. The second component had heavy

loading to blocks and rebounds and was named the "defensive" factor. The final component had heavy loads on field goal percentage and free throw percentage and the author called this the "stability" factor.

The research of Baghal (2012) examined the feasibility of using structural equation modeling (SEM) for multivariate analyses of NBA data. Before the SEM method was conducted, confirmatory factor analysis (CFA) was used to find the offensive Four Factors (effective field goal percentage (EFG), free throw rate (FTR), turnovers per possession (TPP), and offensive rebounding percentage (ORP)) that were indicators of a single latent factor named "offensive quality" and another latent factor named "defensive quality". The SEM was utilized to regress winning percentage on latent offensive and defensive quality as well as salary. It revealed that team salaries were more related to offensive quality instead of defensive quality or winning. The offensive performance also played an important role in winning percentage. It concluded that the money spent was for offensive performance and the offensive quality paid for affected the winning.

Canonical Correlation Analysis (CCA) is most appropriate method when a researcher desires to explore the relationship between two variable sets, since CCA can be easily understood as a method closely linked with Pearson correlation. There are several advantages with CCA. First, it limits the probability of committing Type I error anywhere within the study. Second, it may best honor the reality of psychological research. Lastly, this technique can be used instead of other parametric tests in many instances, making it not only an important technique to learn but a comprehensive technique as well (Sherry, A., & Henson, R. K.,2005).

Cluster analysis is a technique that divides data into clusters (groups) such that the objects within a group have greater similarity with each other and more dissimilarity with objects in other groups (Xiong et al., 2009). Lutz (n.d.) used cluster analysis on NBA player statistics from the 2010 to 2011 season. Herein 10 clusters were found and linear discriminant analysis was used predict cluster membership for the 2008-2009 and 2009-2010 season. This had a misclassification rate of 16%. Lutz's analysis informed the criteria for cluster analysis in this report.

When the ordinary least squares (OLS) method is insufficient for the analysis, a suitable alternative is penalized regression (Ridge regression). Insufficiency of the OLS Model can arise when the number of variables are close to the number of samples which results in instability of the beta coefficients. In addition, insufficiency is seen when the variables are too highly correlated or when the training set is overfitting. To address these limitations, Ridge regression adds a constraint (penalty parameter) to the linear model (James et al. 2014; Bruce and Bruce, 2017) and shrinks the regression coefficients, so that variables with minor contribution to the outcome, have their coefficients close to zero. The amount of the penalty can be fine-tuned using a constant called lambda (λ). Selecting a good value for λ is critical. When $\lambda=0$, the penalty term has no effect, and ridge regression will produce the classical least square coefficients. However, as λ increases to infinite, the impact of the shrinkage penalty grows, and the ridge regression coefficients will get close zero (Bruce, Peter, and Andrew Bruce. 2017).

These studies inspired the use of such five methods on the NBA players dataset. We extended the analysis on more recent data (the 2016-2017 season) compared to the dataset in the research (the 1995-1996 and 2008-2009 seasons). Also, instead of using the team data, we analyzed individual players which could provide more direct evaluation on the performance for each player.

Methods

Principal Component Analysis (PCA): Principal Component Analysis was used to group linearly correlated dimensions into separate principal components and allows us to reduce the dimensionality and discover latent variables. As the game has evolved, players' positions may be less relevant than the style of their play. PCA can potentially allow us to group players not by their official position, but by the style of how they play the game. Orthogonal rotation is done for variable with low correlations and oblique is done for those variables that are highly correlated. Promax is a type of oblique rotation that was applied to the NBA data since many features were highly correlated.

Common Factor Analysis (CFA) and Structural Equation Modeling (SEM): CFA and SEM are statistical techniques utilized to reduce the number of observed variables into a smaller number of latent variables by examining the covariation among them. Latent factors that are highly correlated with the observed variables are extracted from correlation matrix and factor rotation is conducted to increase interpretability. CFA was conducted based on the theoretical models mentioned in the literature review with a mixture of exploratory and confirmatory factor analysis. Based on this research, specific variables were chosen that could reveal two factors (offensive factor and defensive factor). Thus, CFA was used to test whether the latent factors from the observed variables were consistent with past research and understanding of the nature of factors.

The adequacy of sampling was tested through Kaiser Meyer Olkin (KMO). The sampling is sufficient if the value of KMO is larger than 0.5, and preferably between 0.7 and 0.8 (Hutcheson & Sofroniou, 1999). The strength of relationship was tested through Bartlett's test of Sphericity. The null hypothesis of this test was that the correlation matrix is an identity matrix. If the p-value of the test was smaller than 0.05 then the null hypothesis was rejected. For factor rotation, promax rotation, which assumes correlation between the extracted factors, was used as for each player there should be intercorrelation among the latent factor. Relying only on the outcomes of orthogonal rotation could result in the loss of valuable information. Reliability of the factors was assessed using Cronbach's Alpha. Herein, the value was to be larger than 0.70 to indicate that the group of factors were consistent.

CFA is used to confirm and find the latent factors and their related variables (measurement model). Structural Equation Modeling (SEM) was used to find whether relationships exist between variables of interests and factors (structural model). Path model, which shows how the variables of interest related, was constructed to estimate the latent relationship in the underlying process. The focus of SEM was estimating relationships among latent constructs and the directionality of significant relationships. Through SEM, factors of game performance were able to be linked to wining and salary. K-means clustering on factor scores was then utilized as a confirmatory process to segment the players based on the latent factors.

Canonical Correlation Analysis (CCA): CCA was used to identify the strength of association between two sets of variates. Each variate is a linear combination and the coefficients of each linear combination are decided based on maximizing the correlation between canonical variates. Wilk's test was conducted to measure the significance of canonical variates. The canonical analysis was conducted to explore the association between players' skills set and performance. Player's skills set and player's performance measures are identified in **Appendix CCA-1**. The dataset was divided into 3 groups based on players' salary for the 2016-2017 season: High Salary Players who earn more than 10 million dollars, Median Salary Players who earn between 1 ~ 9 million dollars and, Low Salary Players who earn less than 1 million dollars for a year. CCA was conducted on these three group and the results were analyzed to discover interesting patterns.

Cluster analysis was to explore how NBA players could be segmented into groups based on their similarities in indicators of performance within certain players and differences from others. Based on the analysis by Lutz (n.d)., the NBA dataset was reduced by only focusing on players who played at least 30 games during the system and averaged at least 10 minutes per game. Cluster Analysis was applied to the variables listed in **Appendix CA-1**. To segment the players into clusters, both K-means clustering and K-medoids PAM clustering were applied. By applying both techniques, this allowed for a comparison of the most accurate cluster using domain knowledge. Also, K-medoids is known to be more robust and could work more accurately with stellar top NBA players, that may be seen as outliers from an analytic perspective but truly set the tone for performance. To determine the number of clusters, the silhouette method was applied. For visualization, principal component analysis was applied on the variables. The first two principal components were extracted to plot the cluster analysis on these axes.

For the application of multiple linear regression and ridge regression, the data was processed to remove the nominal, ordinal variables and outliers. Linear regression was therefore conducted using the pre-processed data consists of 443 observations and 30 variables out of the 446*38. Feature selection and manual selection were helped to reduce the multicollinearity. After that, the data was split into 70% training and 30% test. The research questions were approached using two different dependent variables: "WINS_RPM" and "Points". Exploratory analysis indicated that the points scored per game have a direct impact on the players' ranking and serve as an indicator of player performance.

The correlation plot (**Appendix RR-1**) identified many independent variables with high correlations that closer to 0.7, indicating a multicollinearity issue. In order to prevent multicollinearity, highly-correlated variables were removed. Both backwards selection and manual selection were used to identify variables with an appropriate VIF as well as to compare the models. For multiple linear regression, 5 folds cross-validation method was used. Ridge regression was also applied to the same dataset to address the multicollinearity issue. For ridge regression, alpha is set to 0. After applying the ridge regression, the best ridge model was found by using the best value of alpha and lambda. Both Ridge and Linear Regression models were compared to find the best model.

Results

Principal Component Analysis

For PCA, the dataset was reduced only include those players that played in 21 or more games (more than 25% of the season). This was to remove any outlier players that maybe only played in one or a few games and had limited statistical reference.

After factor rotation using promax rotation, the number of dimensions was reduced to three principal components that explained 64% of the variance. The scree plot **(Appendix PCA-1)** indicated that the maximum number of optimal principal components was six. However, the knee of the plot was near three. After several iterations, it was determined that three principal components were the optimal selection. The first component explained 32.5% of the variance, the second 16.6%, and the third explained 15.7% for a total 64%. The three components stand out in their significance as well as their independence. **Appendix PCA-2** shows the PCA loadings.

The first component has attributes with positive loadings to two pointers, free throws, assists, steals, turnovers and personal fouls. This component could be labeled "active" for players who are involved in a lot of plays on both defense and offense. This component can also be considered "aggressors" or "attackers." These players are trying to make things happen on the court. They are also taking a lot of free throws because they are forcing the other team to foul them.

The second component, labeled "shooter," has positive loadings to three points attributes and a positive loading to free throw shooting percentage. These players aren't as active but tend to score when they do take shots. Free throw percentage shows up here. Active players take more free throws, but shooters have a higher percentage made. This also has negative loadings to offensive rebounds, blocked shots, defensive rankings. These are variables that would relate to players that play near the basket. Historically, these have been players that are better defensively, but don't take many outside shots. It makes sense to have a positive load to long range shooting and a negative load to rebounds and blocked shots.

The third attribute, labeled "consistent," has positive loadings to two points percentage, effective field goals, games played, average of defensive and offensive ranking, and wins associated to player performance. These players are not necessarily dominant in any one aspect of the game but are above average in many aspects. These players do make a high percentage of the shots they take and have a higher average defensive and offensive ranking. Additionally, these players also play in more games and "show up" every night.

There are three applications applied using the PCA results. **Appendix PCA-3** reflects the visualizations of the first application. First was to look at Individual player component scores and how those scores related to wins. The objective is to seek out players with high "Active" and "Consistent" score as those had a positive relationship to wins. Secondly, strong defensive players also had a positive effect on wins. A derivative for NBA execs is to look for combinations of players that lead to high scores in these components. If you cannot sign Lebron James, there may be a combination of players that have similar attributes.

A second application deals with a Statistical Diversity Index (SDI) (**Appendix PCA-4**). To calculate, a player component scores are compared to all other player scores. The differences in scores make up the SDI. The practical application of an SDI is to find comparable players. For example, the table shows all players with an SDI less than 3 when compared to Harrison Barnes. Barnes and Jabari Parker are highlighted. Barnes and Parker play the same position, have similar component scores and Parker is three years younger. Yet Parker makes $17 million less than Barnes. If Barnes leaves or gets injured, a similar player may be a suitable replacement with the bonus of lower salary.

Common Factor Analysis (CFA) and Structural Equation Modeling (SEM)

Factor analysis focused on the technical court performance and confirmed the two factors theory. There were 11 variables participating in the factor analysis. These variables are identified and defined in **Appendix CFA-1**. To mitigate the impacts of outliers and noises from the data, CFA only focused on players who played more than 30 games and played for their team for at least 10 minutes (N=345).

Suitability of data for Factor Analysis: From the test statistics (**Appendix CFA-2**), the overall KMO statistics were 0.76 which is considered great sample adequacy. From the Bartlett Test of Sphericity, the p-value was less than 0.05 and therefore, this data did not produce an identity matrix and was approximately multivariate normal and acceptable for further analysis.

Factor Extraction: The research of Baghal (2012) used 2 latent factors: Offensive Quality factor and Defensive Quality factor. From the scree plot, 2 factors were appropriate (see **Appendix CFA-3**) with total variance of 70% explained. Finally, 2 factors (63% variance explained) and a 0.51 cut-off were chosen to prevent cross-loading issue.

Factors Labeling: From **Appendix CFA-4A** factor loadings, it is clear that the first common factor has larger loading in the aspects of scoring, assist, steal, offensive ranking which can be named as *attack factors*. The second factor has large loading in the aspects of rebound, block shot, personal foul, defensive ranking and thus can be named as defense factors. It roughly confirmed the results of Baghal (2012); each factor significance is relative reasonable. (See **Appendix CFA-4B** for detailed variables definition and loadings)

Most of the variables under the latent factors had obvious relationships based on the definition of the performance statistics. Steal (STL) might considered as a defensive technique traditionally but actually mainly used by players in the offensive positions (**Appendix CFA-5A**). Offensive rebound might not be a traditional defensive measurement of a player. However, practically, a player with high offensive rebound tends to have high defensive rebound (DRB) and an offensive rebound (ORB) is harder to achieve due to the nature of the game. The strong correlation between ORB and DRB (0.75) (**Appendix-5B)** also confirms the inseparable relationship between ORB and DRB that has the second largest loading for Defense factor. Players in the Center position are more likely to have both high ORB and DRB due to the fact that they are normally the highest player in the team with excellent durability and defensive skills **(Appendix CFA-5C)**. Thus, in our model, ORB has the highest loading in defense factor with a reasonable practical reason.

Internal Consistency: Cronbach's alpha (α) was used to calculate the internal consistency. Both groups have their overall alphas larger than 0.76, which concluded that both groups of factors are internal consistent (see **Appendix CFA-6**).

Structural Equation Modeling

Before constructing the path model, a scatterplot (**Appendix CFA-7**) where the players are colored based on its ranking in the dataset revealed the possible relationship between latent factors, wins and rank. Top 100 players were considered as NBA stars which was reasonable after comparing with the lists of NBA stars voted by fans. Most of the top rank players are high in attack ability. Russell Westbrook and James Harden are outstanding players with high attack abilities. Hassan Whiteside is the top defense player among all the players in the dataset, for he led the league in rebounds and blocks. However, the wins contributed by Russell Whiteside was only 6.28 while Russell Westbrook and James Harden were 17.34 and 15.54, respectively. Structural Equation Modeling is used to quantify the relationship.

In regard to the structural portion of the model (**Appendix CFA-8**), there were clear effects of attack and defensive latent factors on number of wins contributed by the player (WINS_RPM). Further, wins contributed had a positive relationship with salary (scaled due to the large magnitude compared to other variables).

Direct Effects: Attack factor was related positively to wins (standardized coefficient = 0.79) and has higher effect on wins compared to defense factor (standardized coefficient =0.53). Wins is predicted of greater salary (standardized coefficient = 0.52). **Indirect Effects:** attack factor and defense factor might have indirect effect on salary through number of wins contributed.

From the whole SEM process, Attack factor was observed to be more important than the Defense factor. The offensive performance of a player contributes to more winning and higher salary to pay for their excellence in offense.

Clustering by factor scores

To look deeper on the indirect relationship, a more specific research question was raised: Can players be segmented into groups based on the similarities in latent factor scores? To answer this question, factor scores was used to do the k-means cluster analysis as a confirmatory process after CFA. In this process, k-means clustering and k-medoid clustering has a similar result. (**Appendix CFA-9**).

Three clusters gave good results in terms of within-cluster sum of squares (1.96, 3.13 and 3.15 for cluster1, 2 and 3) (see **Appendix CFA-9).** For cluster 1, the cluster mean was high on attack factor scores so it was named Top Attack Players. For cluster 2, the highest cluster mean was on defense factor scores, therefore named Top Defensive Players. For cluster 3, it is called Average Players.

After successfully grouping them using factor scores, an obvious difference was observed in the distribution of wins and salaries in different clusters. For Top Attack Players cluster, it had players with much higher salary on average ($14.7 million) and more wins contributed (8.2) compared to Top Defense Players cluster ($8.4 million, 4.3 wins) and Average Players cluster ($5.3 million, 1.6wins).

Conclusion for Factor Analysis: In conclusion, attack ability of a player contributes more to the winning of the team. High salary is paid for the excellence of a player's attack ability. For a team, it should adjust its resources to focus on strengthening the attack abilities of the whole team. Finding a balance of scoring, attacking, rebounding and attacking is necessary in more winning in the game.

Canonical Correlation Analysis (CCA)

The objective of CCA was two-fold. First, to determine the relationship between player's skills set and player's performance (see **Appendix CCA-1**) and second, to compare the three groups of canonical variates based on three ranges of salary. According to the Wilk's test, for both the high and median salary group, the first three variate appeared to be significant at the 0.05 significance level. On the other hand, only the first two variates were significant for the low salary group based on the chosen significance level. The first canonical variate set of each group explained most of the variance (on average 40%), the rest of this analysis will be based on the first variate sets. The canonical correlations were 0.9854, 0.9778 and 0.9695 which demonstrated a strong positive correlation between two variates (see **Appendix CCA-2**).

Of player's skills set, in high salary group (see **Appendix CCA-3A**), Turnover weighed the most but in a negative way, then 3-Points Field Goals weighed the second and Steal weighed the third. The coefficient of Block and Personal Foul is small enough to be omitted. Other skills weighed even. Indicates that mistakes are critical among high salary player since usually they play as leading role in the game, one mistake happens on them might change the game. In median salary group (see **Appendix CCA-3B**), the three most important variables look the same, however Block is important to them and Turnover has less impact on them comparing to high salary group. In low salary group (see **Appendix CCA-3C**), 3-Points Field Goals, Steal and Block weighed the most. 3-Point Field Goals obviously weighed more and Turnover on the contrary has less impact comparing to high and median salary group, so scoring skill looks more important to low salary player and mistakes tend to have less impact on them. Overall 3-Points Field Goals is the suppressor variable of player's skill set which indicates how important is scoring skill to a player when measuring their performance. Of player's performance, Player Impact Estimate is the most important variable when measuring player's skill and Min Played per Game weights the second in common which means when measuring a player's skill or ability, how they actually contribute to the game not only relies on how many scores they can get but also considering their scoring efficiency and other factors. Min Played per Game reflects a player's importance and contribution to the game since a player performs better the longer time he plays. However, Game Played can't really provide useful information of a player's ability as even though a player plays a lot of games

but most of the time he contributes to the game is when the result has been decided and he can't change the game a lot with his ability so the coefficient of Game Played is small enough to be omitted.

Cluster Analysis

K-means: From the silhouette plot, the optimal number of clusters was three. **Appendix CA-2A** shows the cluster means based on the k-means algorithm. These clusters could be labeled as:

1) *Offensive rebounders and good shooters:* these players are strong in getting close to the basket for offensive rebounds and are consistent in making free throws when fouled.
2) *High scorers:* These individuals are dominate in averaging the most points during the game and have strong offensive performance.
3) *Poor defenders:* These individuals are have strong performance in defensive rebounds but concurrently possess a high average of personal fouls. As both their rebounds and fouls are important, they are not efficient defenders.

K-medoid PAM: Based on the silhouette plot, three clusters were chosen. **Appendix CA-2B** shows the medoids of the three clusters. These clusters could be labeled as

1) *High scorers:* These are mainly the top NBA players that are high scorers.
2) *Excellent defenders:* These players are great at getting close to the basket and are thus high in both offensive and defensive rebounds.
3) *Great attendance:* These players are constant participants in games although they may not be strong in offensive or defensive performance.

PCA: PCA with varimax rotation was conducted. PCA summary is displayed in **Appendix CA-3**. The first two components were to be used as axes on the plot to help interpret the cluster results. These two components explained 60.3% of the variance. PC1's most important variable was points whereas for PC2 it was offensive rebounds.

Differences and implications: While the results of k-means clustering and PAM clustering were similar, there were some noticeable differences. K-means highlighted poor defenders which is an important implication for team managers. A player high in defensive rebounds but also high in personal fouls can be a liability in that many personal fouls can have the player removed from the game as well as increase the number of team fouls leading to free throw attempts by the opposing team or a loss of possession. Another difference was the clustering of Anthony Davis one of the top defenders in the NBA. With K-means, he was in the same cluster as the other players, which made sense according to his ranking. Whereas PAM clustering, indicated that Anthony Davis was best fit for cluster 2, which made sense given his high defensive performance. This is displayed in **Appendix CA-4**.

Wins in basketball requires both offensive and defensive skills. Therefore, for practical application PAM clustering provided the best clusters. Interestingly, it highlighted great attendance indicating players that are not highly skilled in either area but are consistent and contributed to games. This would be an area for future analysis to determine the value derived from playing these individuals.

Further look at players' positions: The next objective within cluster analysis was to identify the position within each cluster. **Appendix CA-5A** shows that centers make up the majority of cluster 2. Given their height, and position on the court, this makes practical sense as they play close to the basketball and are therefore in the best position to be great defenders, catching rebounds both offensively and defensively. The distribution of the other players are provided in **Appendix CA-5B**. Cluster 1 is dominated by the Top NBA players that are mainly point guards, small forwards and shooting guards. These players are

identified in **Appendix CA-5C**. By contrast, cluster 3 is comprised of these players along with power forwards that are weak in both offensive and defensive performance.

Ridge Regression

For Estimate of number of wins a player has contributed (WINS_RPM) ***note on the same analysis for points

By applying linear regression on training data and got the results shown in **Appendix RR-2**. Ridge regression helped to shrink coefficients by keeping all variables in the model.

The multiple linear regression results are shown in **Appendix RR-3A**. **Appendix RR-3B** presents the RMSE, R-Squared and MAE values for both using backward selection and manual selection. The RMSE (root mean square error) for WINS_RPM is 0.89 and r squared is 0.93 which indicates that almost 93% of variability seen in response (WINS_RPM) is explained by backward selection model. The MAE (mean absolute error) is 0.69 for backward selection. The RMSE (root mean square error) for WINS_RPM is 0.998 and R-squared is 0.92 which indicates that almost 92% of variability seen in response ("WINS_RPM") is explained by the manual selection model. The MAE is 0.756 for manual model.

The residual plots are shown in **Appendix RR-4**. **Appendix RR-5** Ridge regression-plot shows the results of ridge regression. On Y-axis, the root means square error which has been estimated using repeated cross-validation. On X-axis is the lambda, and the RMSE values increase as the lambda increases. The minimum RMSE generated when the lambda value is 0.26. **Appendix RR-6** Ridge Regression -log lambda plot shows the log lambda value on X-axis and coefficients on Y-axis so when log lambda is about 8, all the coefficients are more or less zero and as lambda is released coefficients starts to grow. As coefficients start to grow, the sum of the square of coefficients becoming larger. The figures show that all 19 variables on each point for backward selection model and 17 for manual selection model. **Appendix RR-7A** shows the plot of variable importance. **Appendix RR-7B** shows that the most key variables that impact the estimate of numbers of wins a player has contributed are eFG., STL, PF, BLK, DRM and ORM for backward selection model; STL, RPM, X2P, PF, BLK, and ORM for manual selection model. The equations for linear regression:

*Wins_PRM (Backward) = 6.12 - 0.02*Age - 0.07*MP - 0.9*X3P – 0.51*X2PA +1.45*eFG% + 0.41*ORB + 0.45*DRB + 0.63*AST + 0.74*STL + 0.39*BLK – 0.84*TOV - 0.82*PF + 0.65*POINTS + 0.79*ORPM + 0.97DRPM - 0.28*PIE + 0.03PACE*

*Wins_PRM (Manual) = 3.93 - 0.03*Age + 0.47 *X3P – 0.51*X3% - 0.86*X2% +0.40*FTA + 0.12*FT% + 0.59*ORB + 0.26*AST + 0.49*STL + 0.62*BLK - 0.74*PF + 1.01*RPM - 0.17*PIE*

The variables with underlines are shared variables from both backward selection and manual selection which have impact on the estimate of number of wins a player has contributed (WINS_RPM).

Appendix RR-8 shows the comparison between the linear regression and ridge regression results. It gives a summary in the form of min, 1st quartile, median, mean, 3rd quartile and maximum. The mean values of RMSE for both the models are relatively close. As the differences are very small from both models, it is difficult to find Ridge Regression model is more significant. **Appendix RR-9** shows that the best model of the ridge is at alpha is zero and lambda is 0.26.

Points

The same analysis process was applied to predict POINTS from the same dataset used with WINS_RPM. For brevity, these results are in **Appendix RR-10.**

Limitations and Future Work:

Different methods of multivariate analysis were applied to see if some interesting insights can be uncovered around player rankings, salary, games won, or any number of other potential outcomes. There may be some player attributes that team executive value higher than others which may lead to a higher salary or wins for a player. This CFA only involves two factors from 11 variables to confirm the model proposed by past research paper. There might be more latent variables and variables can be added for analysis to improve the model. From the CFA and SEM, the models may be helpful to acquire talent to fill necessary role but it would be better if including some features to identify the non-measurable qualities of a players to fully evaluate the potential of a player in the future.

In the future, the cluster analysis could be followed by linear discriminant analysis to predict cluster membership of players in more recent seasons. It would also be of value to analyze player performance over time and additionally by team to see if certain teams enhance or inhibit performance along certain statistics. Lastly, creating profiles of the most important performance measures would be valuable to inform NBA draft and trades between teams.

Conclusion

The initial dataset had 446 instances with 38 features. Some of these variables were highly correlated so we used multivariate techniques learned in this class including PCA, CFA, CCA, Cluster Analysis and Ridge Regression. It was evident that we could move beyond the traditional classification of players by position types that were assigned based on a player's physical attributes. All classification techniques were able to distinguish players and assign attributes on style of play rather than position played. PCA, CFA and Cluster analysis were able to assign these players on two or three distinct styles of play. CCA furthered the analysis and was able to show that different skill sets were more impactful on performance. Finally Ridge Regression also showed how different features impacted performance. We believe this analysis can have significant impacts on the way management build teams in the future. No longer will they have to look to fill a Center position or a Guard position. They can look to fill with an attacker, rebounder, or shooter and maximize wins on those attributes.

References

Baghal, Tarek. (2012). Are the Four Factors Indicators of One Factor? An Application of Structural Equation Modeling Methodology to NBA Data in Prediction of Winning Percentage. Journal of Quantitative Analysis in Sports, 8(1). doi:10.1515/1559-0410.1355.

Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists*. O'Reilly Media.

Cheng, A. (2017, March 2). Using Machine Learning to Find the 8 Types of Players in the NBA. *Fastbreak Data*. Retrieved from https://fastbreakdata.com/classifying-the-modern-nba-player-with-machine-learning-539da03bb824

Hutcheason, G. D. and Sofroniou, N. (1999). The Multivariate Social Scientist: an Introduction to Generalized Linear Models. Sage Publications.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Lutz, D. (n.d.). Cluster analysis of NBA players. Sloans Sport Conference. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2012/02/44-Lutz_cluster_analysis_NBA.pdf

Xiong H., Wu, J., & Chen, J . (2009). K-means clustering versus validation measures: a data-distribution perspective. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 318-331.

Yin, Wei (2014). Principal Component Factor Analysis-Based NBA Player Comprehensive Ability Evaluation Research. Journal of Chemical and Pharmaceutical Research, 2014, 6(6):2400-2405. Retrieved from **https://pdfs.semanticscholar.org/1bd2/e1928d78b92f96ae767837d650dbf15246fa.pdf**

Appendix

Appendix A (in the order of the paper for each techniques)

Methods	Group Member
PCA	Adam

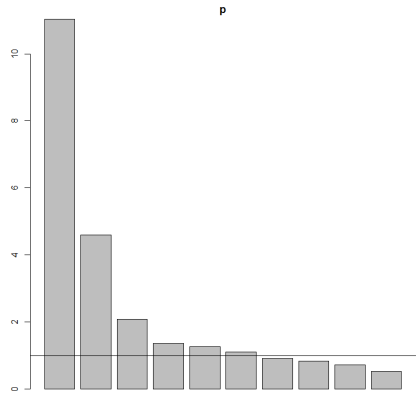
CFA SEM Clustering	Tianyi Tan
CCA	Xiaotong
Clustering	LaShawn
Ridge Regression	You Yue

Appendix B: All the Variables and Definitions

Variable	**Description**
Rk	Rank
MP	Minutes played
FG	Field Goal
FGA	Field Goal Attempts
FG%	Field Goal Percentage; the formula is FG/FGA.
3P	3-Point Field Goals
3PA	3-Point Field Goal Attempts
3P%	3-Point Field Goal Percentage
2P	2-Point Field Goals
2PA	2-Point Field Goal Attempts
2P%	2-Point Field Goal Percentage; the formula is 2P/2PA
eFG%	Effective Field Goal %
FT	Free Throws
FTA	Free Throws Attempts
FT%	Free Throws Percentage; the formula is FT/FTA.
ORB	Offensive Rebounds
DRB	Defensive Rebounds
TRB	Total Rebounds

AST	Number of Assists
STL	Number of Steals
BLK	Number of Blocks
TOV	Turnovers: plays when player loses ball to defense
PF	Number of Personal Fouls
POINTS	Points scored per game
TEAM	Team
GP	Number of games played
MPG	Minutes per game played (this is the same as MP so not sure why there are two)
ORPM	Players impact on team's offensive performance
DRPM	Players impact on team's defensive performance
RPM	Real Plus-Minus
WINS_RPM	Estimate of number of wins a player has contributed. The team won this more games because of this player.
PIE	Player Impact Factor: player's contribution vs. statistics in games they play
PACE	Pace Factor: the number of possessions per 48 minutes
W	Games played where the team won.

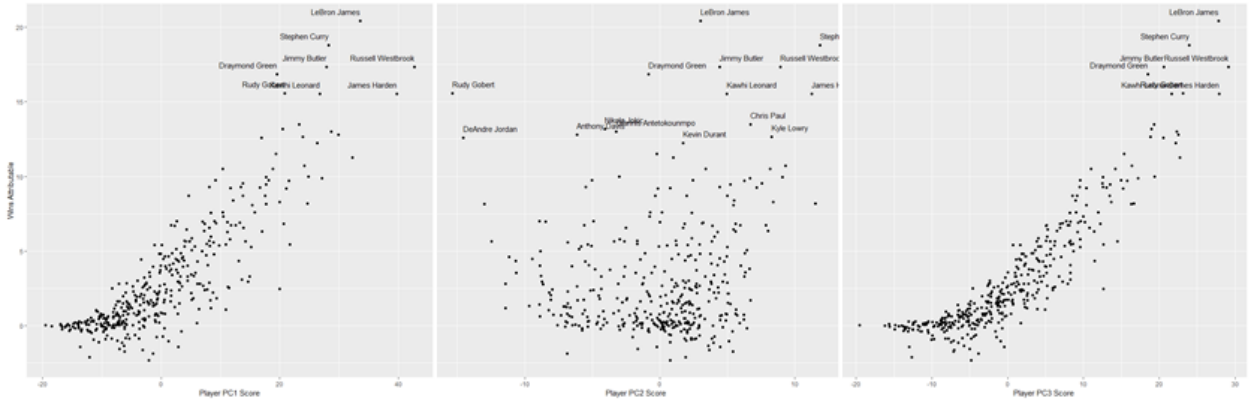
Appendix (Principal Component Analysis)
PCA-1: scree plot



PCA-2: PCA Loadings

```
Loadings:
          RC1    RC2    RC3
MP         0.727
X2P        0.965
X2PA       1.031
FT         0.971
FTA        0.992
DRB        0.637
AST        0.737
STL        0.620
TOV        0.996
PF         0.536
PIE        0.704
X3P                0.763
X3PA               0.762
X3P.               0.758
FT.                0.593
ORB               -0.767
BLK               -0.634
DRPM              -0.535
X2P.                      0.639
eFG.                      0.866
GP                        0.575
RPM                       0.709
WINS_RPM                  0.566
W                         0.827
AGE
ORPM
PACE

                 RC1   RC2   RC3
SS loadings    8.769 4.486 4.247
Proportion Var 0.325 0.166 0.157
Cumulative Var 0.325 0.491 0.648
```

PCA-3: Visualization of PCA Application



PCA-4: Statistical Diversity Index

PLAYER	RC1	RC2	RC3	SDI	SALARY	POSITION	AGE
Harrison Barnes	8.9034	1.3565	5.5778	0	22,116,750	PF	24
Elfrid Payton	9.7602	0.5276	6.1368	1.3167	2,613,600	PG	22
Tobias Harris	8.093	1.7144	6.8998	1.5913	17,200,000	PF	24
Tyler Johnson	7.6164	2.2981	5.1142	1.6607	5,628,000	PG	24
Wilson Chandler	7.0378	1.2937	5.2251	1.8997	11,233,146	SF	29
Ersan Ilyasova	7.3287	1.8456	6.9242	2.1289	8,400,000	PF	29
Victor Oladipo	9.2295	3.1172	7.2725	2.4654	6,552,961	SG	24
Jabari Parker	11.2751	1.0628	6.2289	2.4769	5,374,320	PF	21
Dwyane Wade	11.4364	1.8093	5.969	2.6027	23,200,000	SG	35
Marcus Smart	8.383	4.0293	5.2317	2.7449	3,578,880	SG	22
Patrick Beverley	8.3648	2.3796	8.1948	2.8611	6,000,000	SG	28

Appendix (Factor Analysis)

CFA-1: CFA Variables and Definition

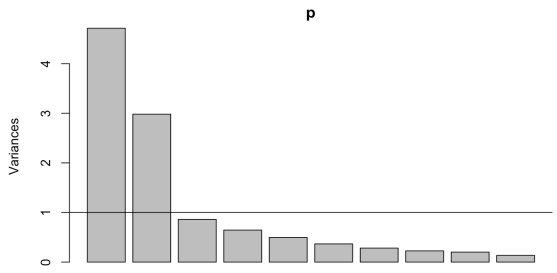
Variable Names	Definition
ORPM	Player's average impact on his team's offensive performance, by the points scored per 100 offensive possessions

FT	Free throw made
AST	Assists (helping somebody else score. If you pass the ball to another player and he shoots and scores, you get an assist)
X3P	3 pointers made
STL	Steals (defensive play where you take the ball from the offensive player before they shoot)
X2P	2 pointers made
ORB	Offensive rebound
DRB	Defensive rebound
BLK	Blocked shot
PF	Personal foul per game
DRPM	Player's average impact on his team's defensive performance, by the points allowed per 100 offensive possessions

CFA-2: KMO and Bartlett's Test

Kaiser Meyer Olkin (KMO)	Overall	0.76
Bartlett Test of Sphericity	Bartlett's K-squared	2020
	Df	10
	P-value	<0.0000000000000002

CFA-3: Total Variance Explained



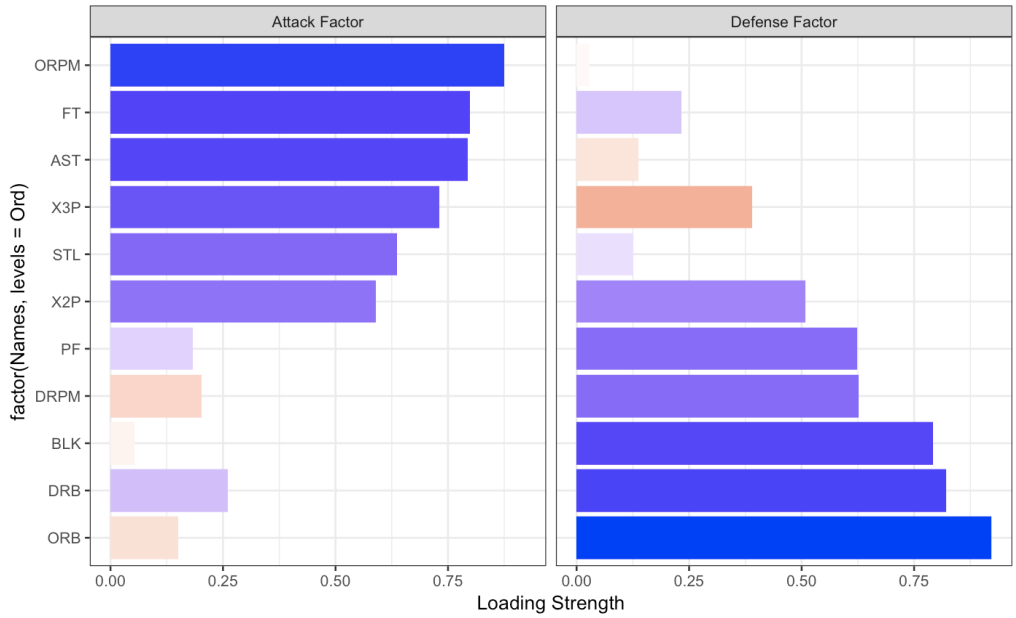
Importance of Components	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	2.171	1.727	0.9279	0.8035	0.7046	0.6055	0.5318	0.4758	0.4493	0.3679	0.29515
Proportion of Variance	0.429	0.271	0.0783	0.0587	0.0451	0.0333	0.0257	0.0206	0.0184	0.0123	0.00792
Cumulative Proportion	0.429	0.700	0.7780	0.8367	0.8818	0.9151	0.9408	0.9614	0.9798	0.9921	1.0000

CFA-4A: Common Factor Analysis Factor Loadings by Latent Factors

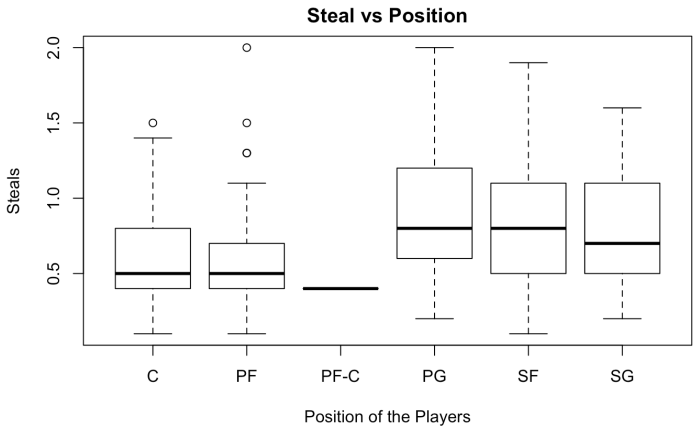
NBA Statistics	Factor Loadings
Factor 1 Attack Factor	
ORPM Player's average impact on his team's offensive performance, by the points scored per 100 offensive possessions	0.875
FT Free throw made	0.799
AST Assists (helping somebody else score. If you pass the ball to another player and he shoots and scores, you get an assist)	0.793
X3P 3 pointers made	0.731
STL Steals (defensive play where you take the ball from the offensive player before they shoot)	0.637
X2P 2 pointers made	0.590
Factor 2 Defense Factor	
ORB Offensive rebound	0.922
DRB Defensive rebound	0.820
BLK Blocked shot	0.791

DRPM Player's average impact on his team's defensive performance, by the points allowed per 100 offensive possessions	0.626
PF Personal foul per game	0.624

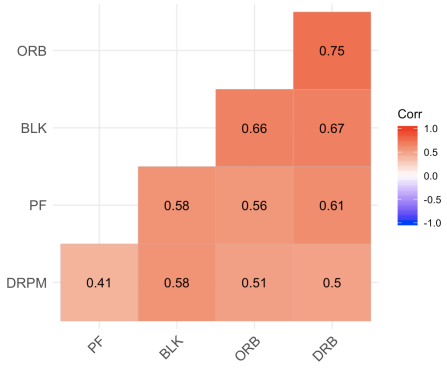
CFA-4A: Common Factor Analysis Factor Loadings by Latent Factors



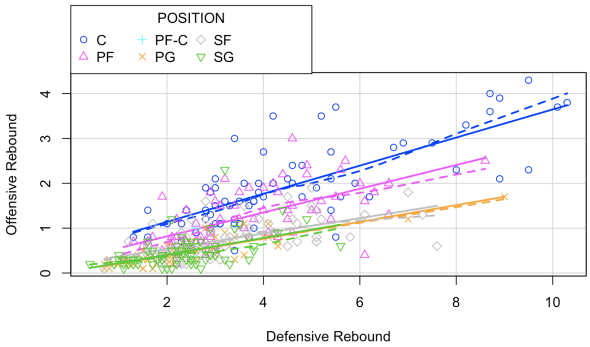
CFA-5A: Boxplot: Steal vs Position



CFA-5B: Correlation Table



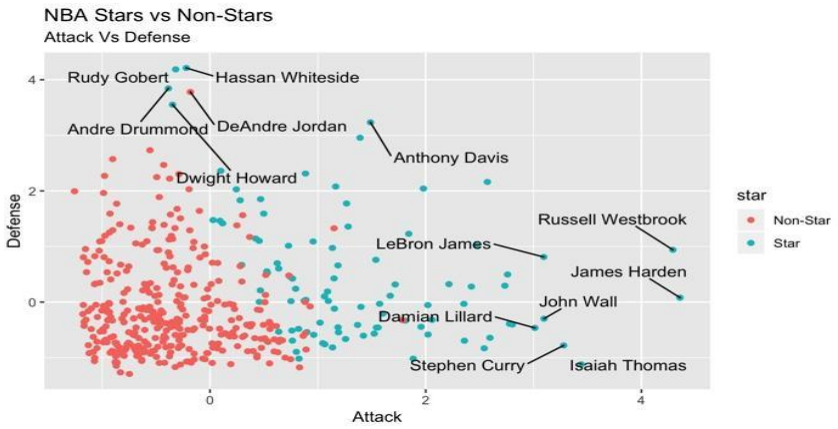
CFA-5C: Scatterplot: Defensive Rebound vs Offensive Rebound



CFA-6: Cronbach's alpha

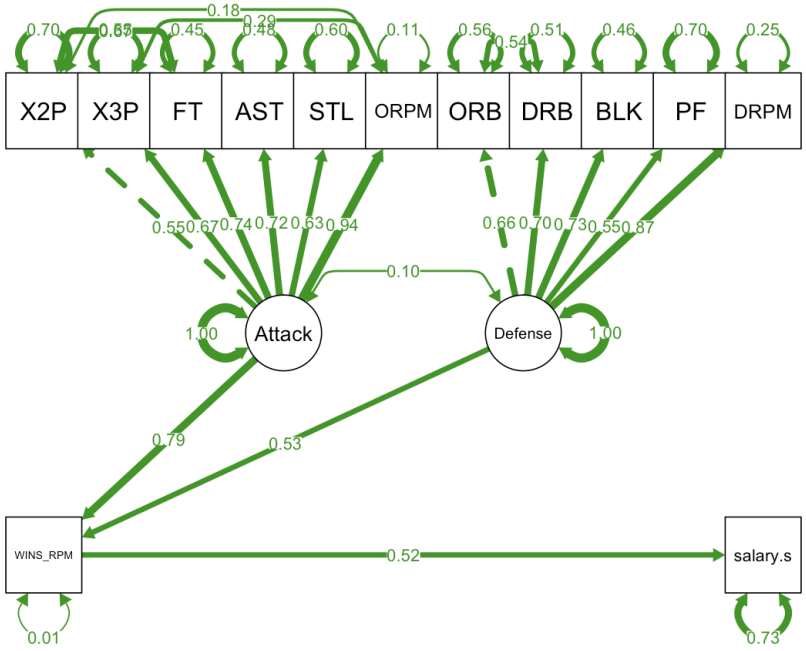
Factor Group	Cronbach's Alpha	[Lower, Upper]
Attack	0.78	[0.75,0.81]
Defense	0.84	[0.82, 0.86]

CFA-7: CFA Scatterplot using Latent Factor Scores



(Star: top 100 players ordered by points per game)

CFA-8: Structural Equation Modeling Path Model Results



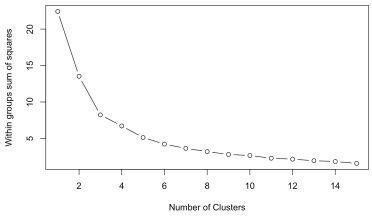
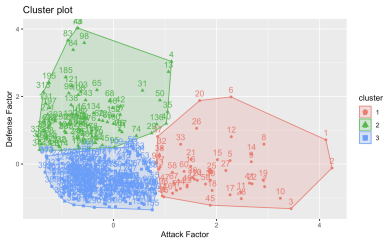
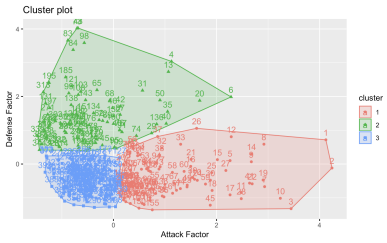
(standardized parameters are shown on the edges, see Appendix Table 7 for detailed statistics)

Regressions	Estimate	Std.Err	z-value	P(>|z|)	Standardized
WINS_RPM					
Attack	3.181	0.277	11.505	0.000	0.791
Defense	3.837	0.273	14.044	0.000	0.582
Salary (scaled)					
WINS_RPM	0.293	0.026	11.219	0.000	0.517

Covariance	Estimate	Std.Err	z-value	P(>|z|)
Attack				
Defense	0.053	0.032	1.639	0.101

Chi Square: 1177 Degree of Freedom: 59 P-value<0.05 GFI: 0.672

CFA-9: Cluster Analysis

Scree Plot for Determination of Number of Clusters	K-means Cluster Plot	K-medoid Cluster Plot
		

Cluster Statistics -1

Cluster	Name	Attack	Defense	SS Within
1	Top Attack Players	0.3300	-0.040	1.9600
2	Top Defense Players	-0.096	0.2500	3.1300
3	Average Players	-0.039	-0.097	3.1500

Cluster Statistics -2

Cluster	Name	Size	Average Salary	Average Wins (Contributed)	NBA Stars (Top 100 rank)
1	Top Attack Players	50	$14,701,180	8.2	98%
2	Top Defense Players	87	$ 8,396,763	4.3	26%
3	Average Players	208	$ 5,307,043	4.6	10%

Appendix (Canonical Correlation Analysis):

CCA-1: Canonical Correlation Analysis variables

N=446	Players' skills set	Performance measures
Number of variables	9	5
Redundancy index on average	0.5557392	0.6165086
Variables: 1	Field Goals	Minutes Played per game
2	2-Point Field Goals	Game played
3	3-Point Field Goals	WINS_RPM
4	Free Throws	Player Impact
5	Offensive Rebounds	Salary
6	Assists	
7	Defensive Rebounds	
8	Steals	
9	Blocks	

CCA-2: Coefficients of the first variate of each group

	High Salary Group	Median Salary Group	Low Salary Group
Player's Performance			
2-Points Field Goals	0.2129	0.2319	0.0801
3-Points Field Goals	0.3548	0.5183	0.7411
Free Throw	0.1297	0.0800	0.2737
Assist	0.1961	0.2208	0.2797
Offensive Rebound	0.1567	0.1706	0.3328
Defensive Rebound	0.1599	0.2072	0.1979
Steal	0.2648	0.5634	0.6745
Block	0.0688	0.1561	0.3592
Personal Foul	-0.0526	-0.0202	-0.0600
Turnover	-0.4346	-0.3759	-0.1883
Player's Value			
Min Played per Game	0.0718	0.0995	0.1214
Game Played	-0.0028	-0.0025	-0.0017
WIN_RPM	0.0346	0.0624	0.0952
Player Impact Estimate	0.1577	0.0884	-0.0655

CCA-3A: High Salary Group variate summary

High Salary Group

Wilk's test

```
     WilksL      F df1     df2     p
[1,]  0.010 20.455  40 346.917 0.000
[2,]  0.349  4.329  27 269.330 0.000
[3,]  0.698  2.287  16 186.000 0.004
[4,]  0.878  1.863   7  94.000 0.084
```

Canonical Correlation

```
Canonical Correlations:
     CV 1      CV 2      CV 3      CV 4
0.9854023 0.7072516 0.4526784 0.3490492
```

Coefficients

```
X Coefficients:
           CV 1        CV 2        CV 3       CV 4
X2P  0.21290982  0.39239589  0.68888326  0.1776686
X3P  0.35489913 -0.65917694  0.67535817  0.9373668
FT   0.12977942 -0.17684134 -0.49220378 -0.3565858
AST  0.19617383 -0.18554921 -0.17068786  0.2546899
ORB  0.15676497 -0.18348564 -0.33792199  1.2542055
DRB  0.15998204  0.09662792 -0.20647806 -0.2241255
STL  0.26488124 -1.12743880  0.30434335 -0.7591607
BLK  0.06882229 -1.02230090 -0.50926679  0.9916590
PF  -0.05260872 -0.05065604  0.88681255 -1.0884138
TOV -0.43469110  0.82499885 -0.06508749 -0.2954083

Y Coefficients:
                CV 1         CV 2         CV 3        CV 4
MPG       0.071848904  0.003459714  0.224633706 -0.07482550
GP       -0.002827944  0.024754641  0.004586787  0.08520779
WINS_RPM  0.034624271 -0.313356913 -0.160281771  0.02288581
PIE       0.157737361  0.356018060 -0.131079203 -0.01535970
```

CCA-3B: Median Salary Group variate summary

Median Salary Group

Wilk's test

```
     WilksL      F df1     df2    p
[1,]  0.016 48.505  40 972.577 0.00
[2,]  0.357 11.766  27 751.215 0.00
[3,]  0.691  6.547  16 516.000 0.00
[4,]  0.975  0.961   7 259.000 0.46
```

Canonical Correlation

```
     CV 1      CV 2      CV 3      CV 4
0.9778820 0.6952228 0.5395267 0.1591044
```

Coefficients

```
X Coefficients:
           CV 1        CV 2       CV 3       CV 4
X2P -0.23195588  0.06918711  0.7779037 -0.4277114
X3P -0.51831313  0.35820354  0.1235053  0.4988851
FT  -0.08000317 -0.59189134 -0.4938445  0.6646347
AST -0.22083711 -0.31174215 -0.3081150 -0.3881560
ORB -0.17068874 -0.74466027 -0.3544690  0.4864125
DRB -0.20729246 -0.26066556 -0.2419372 -0.6225857
STL -0.56346107  0.56204390 -1.3881039  2.2593659
BLK -0.15610820 -0.99214304 -0.6911125  1.0223167
PF   0.02028368  1.44419996 -0.6011796 -1.0082889
TOV  0.37591944  0.77526645  1.7617668  0.8245423

Y Coefficients:
                CV 1        CV 2        CV 3        CV 4
MPG      -0.099567818  0.08143664  0.10218990  0.08250636
GP        0.002513923  0.01337247 -0.01020142 -0.05632528
WINS_RPM -0.062419206 -0.13635922 -0.48180109  0.12290866
PIE      -0.088449015 -0.29138879  0.21703696 -0.11083482
```

CCA-3C: Low Salary Group variate summary

Low Salary Group

Wilk's test

```
     WilksL     F df1     df2     p
[1,]  0.028 8.580  40 217.993 0.000
[2,]  0.462 1.908  27 170.032 0.007
[3,]  0.673 1.615  16 118.000 0.075
[4,]  0.881 1.157   7  60.000 0.341
```

Canonical Correlation

```
Canonical Correlations:
     CV 1      CV 2      CV 3      CV 4
0.9695176 0.5603090 0.4859559 0.3448646
```

Coefficients

```
X Coefficients:
           CV 1       CV 2        CV 3       CV 4
X2P -0.08015242  0.8254407  0.19588219  1.3114008
X3P -0.74119036  0.1528097  0.88530041  0.7202358
FT  -0.27371033 -0.2024189 -0.70073520 -0.6083932
AST -0.27976742 -0.1055265 -0.34953206 -1.0250706
ORB -0.33282245  1.0444981 -1.49817671 -0.6654999
DRB -0.19798953 -0.3629625 -0.25766636 -0.4419455
STL -0.67453414  0.3665254  1.34191401 -0.8099767
BLK -0.35929016  2.6139251  4.26455116 -2.0714866
PF  -0.06004721 -1.9051314  0.16676084  0.6848227
TOV  0.18835084  0.5456684  0.02434312  1.1473714

Y Coefficients:
                CV 1         CV 2        CV 3        CV 4
MPG      -0.121446171 -0.075574392 -0.01002391  0.12923585
GP        0.001731661 -0.001201822  0.04186761 -0.02741473
WINS_RPM -0.095209258 -0.225362671 -0.74793160 -0.68264838
PIE      -0.065537191  0.269618448 -0.02185997  0.00535404
```

Appendix 5 (Cluster Analysis)

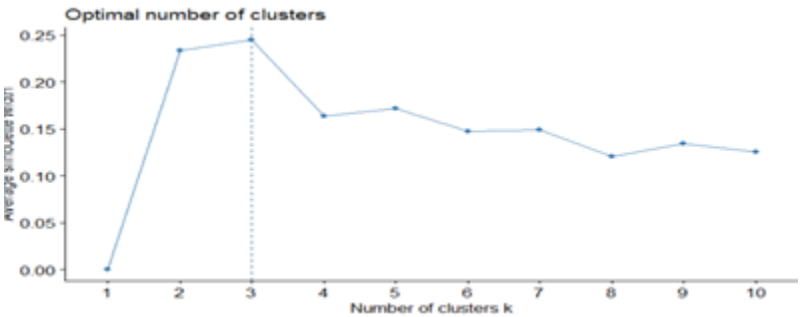
CA-1: clustering variables

The same variables were used for both k-means clustering and k-medoids clustering:

```
> names (nba_kmed)
 [1] "FG."    "X3P."   "X2P."   "FT."    "ORB"    "DRB"    "AST"    "STL"
 [9] "BLK"    "TOV"    "PF"     "POINTS" "GP"     "MPG"    "ORPM"   "DRPM"
```

CA-2A: silhouette plots for K-means

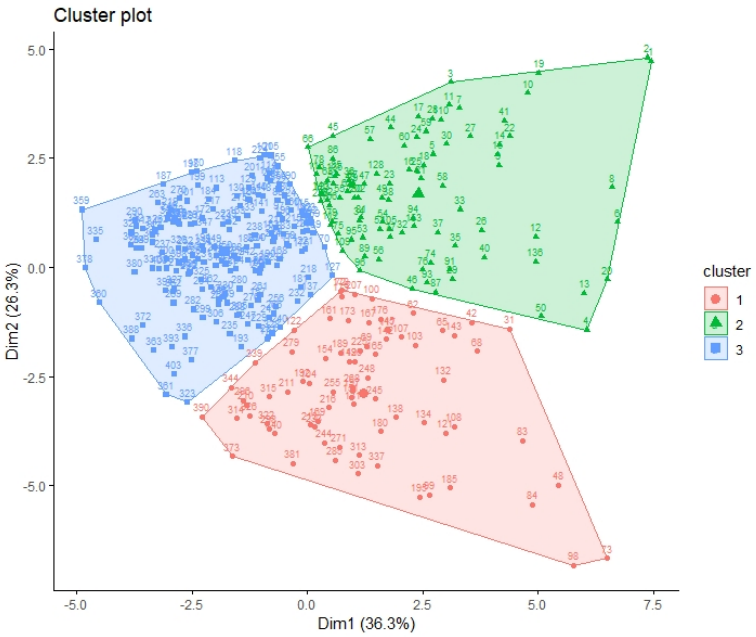
Silhouette Plot of K-means



Summary K-means clusters

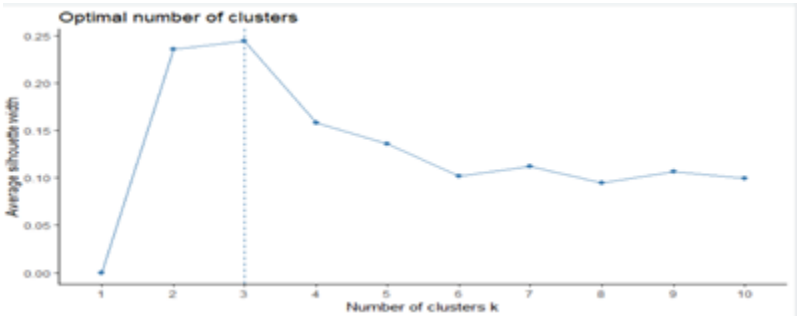
```
> aggregate(nba,by=list(k3$cluster),FUN=mean)
  Group.1         FG.        X3P.         X2P.         FT.         ORB         DRB
1       1  1.23816173 -1.1455362  0.91142337 -0.8846195  1.32056393  0.7446366
2       2 -0.03619078  0.4514032  0.02405187  0.5274084  0.01216653  0.5279897
3       3 -0.51553721  0.2561378 -0.40627841  0.1033074 -0.57667242 -0.6005422
         AST         STL         BLK         TOV         PF      POINTS
1 -0.5496621 -0.1369212  1.01802834 -0.1371571  0.7798078 -0.2073056
2  1.0853204  1.0131143  0.07536233  1.1017569  0.4757229  1.2311719
3 -0.3361391 -0.4762137 -0.47942528 -0.5229514 -0.5881115 -0.5610441
           GP         MPG        ORPM        DRPM
1  0.08214026 -0.08073461 -0.4426515  0.9182962
2  0.37337753  1.17697383  1.0794863 -0.0557849
3 -0.23276574 -0.58706100 -0.3792654 -0.3670598
```

Cluster plot



CA-2B: silhouette plots for K-medoids

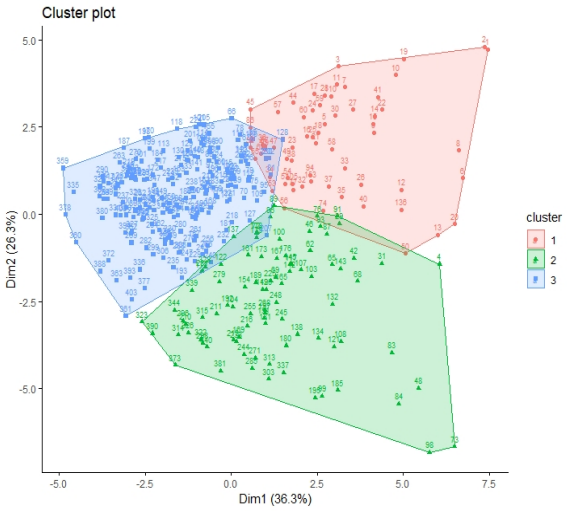
Silhouette Plot of K-medoids



Summary K-medoids

```
> pam_nba$medoids
           FG.        X3P.        X2P.         FT.         ORB         DRB
25   0.2427360  0.7781738  0.1740022  0.8689583 -0.3459415  0.8119031
120  0.9340556 -0.5528187  0.4036907 -0.3976796  1.2635476  0.5310100
238 -0.4014481  0.1485426 -0.3674066  0.2503677 -0.7173620 -0.5363836
           AST         STL         BLK         TOV          PF      POINTS
25   0.67771836  0.6676491 -0.3749841  0.75412913 -0.5418869  1.96751443
120 -0.70264805 -0.6278455  0.7884109 -0.01561806  0.2882684  0.09482056
238  0.09371718  0.1494513 -0.3749841 -0.40049166 -0.5418869 -0.59689519
           GP        MPG        ORPM        DRPM
25   0.5195854  1.5317098  1.7732435 -0.03012385
120  0.8755365  0.2909493 -0.6062925  1.21345030
238  1.0179169 -0.2880722 -0.4314695 -0.50386638
```

Cluster plot

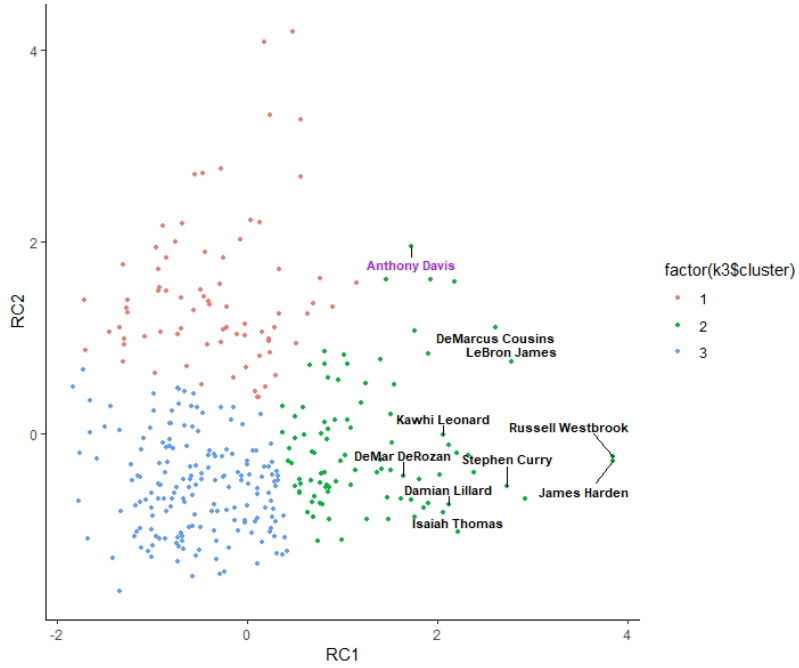


CA-3: PCA

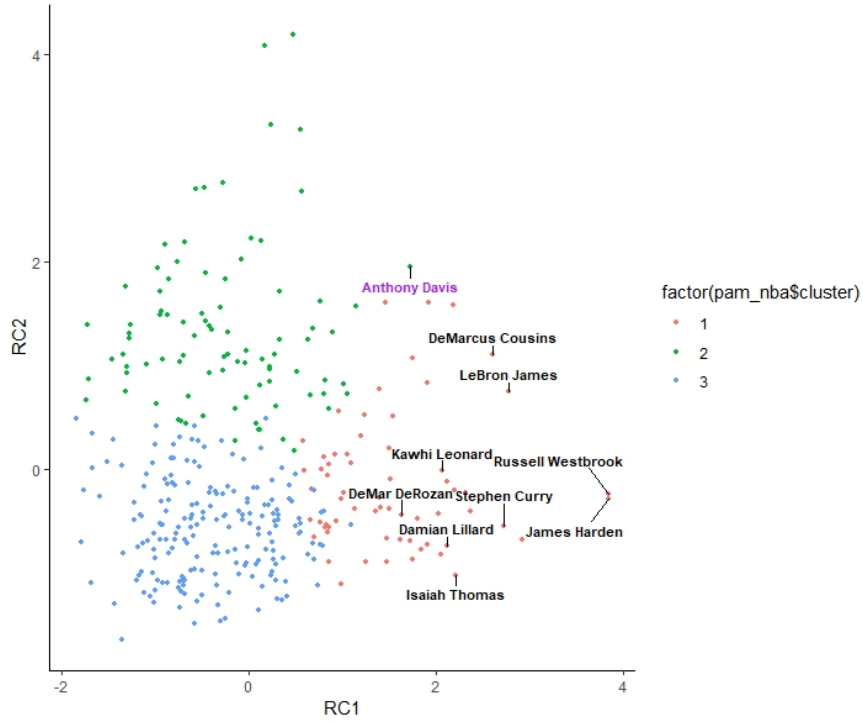
```
> print(nba2$loadings, cutoff=.53, sort=T)

Loadings:
       RC1    RC2
AST     0.803
STL     0.763
TOV     0.872
POINTS  0.909
MPG     0.893
ORPM    0.839
FG.             0.819
X3P.           -0.630
X2P.            0.720
FT.            -0.546
ORB             0.884
DRB             0.698
BLK             0.777
PF              0.597
DRPM            0.681
GP

                 RC1   RC2
SS loadings    5.326 4.689
Proportion Var 0.333 0.293
Cumulative Var 0.333 0.626
```

CA-4: Clustering Plots

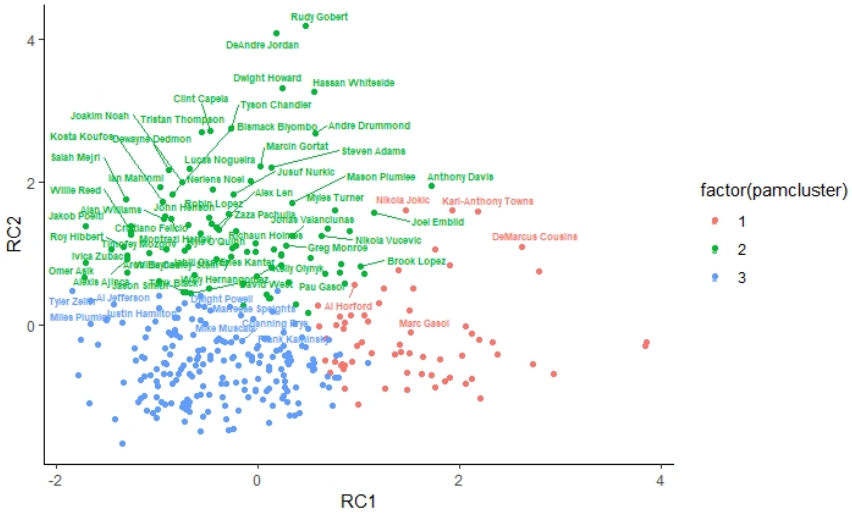
K-means



K-medoids



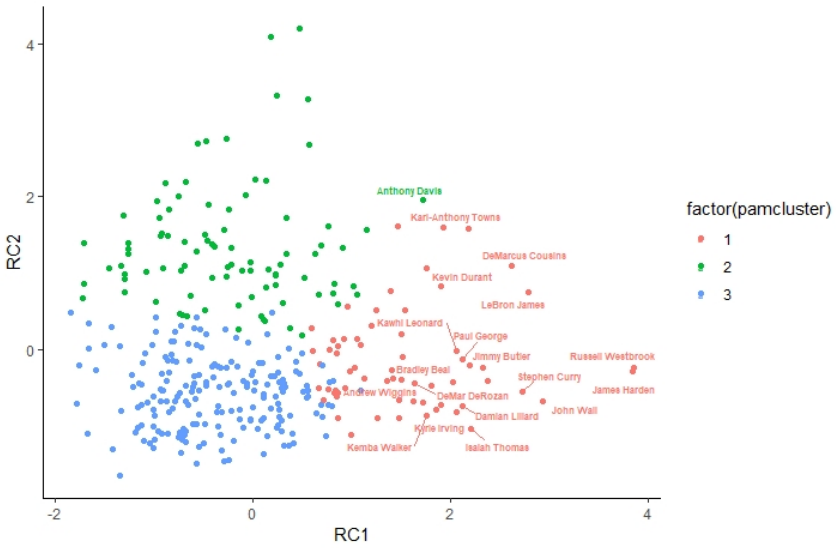
CA-5A: Plots of Centers by PCA components based on PAM results



CA-5B: Positions by cluster

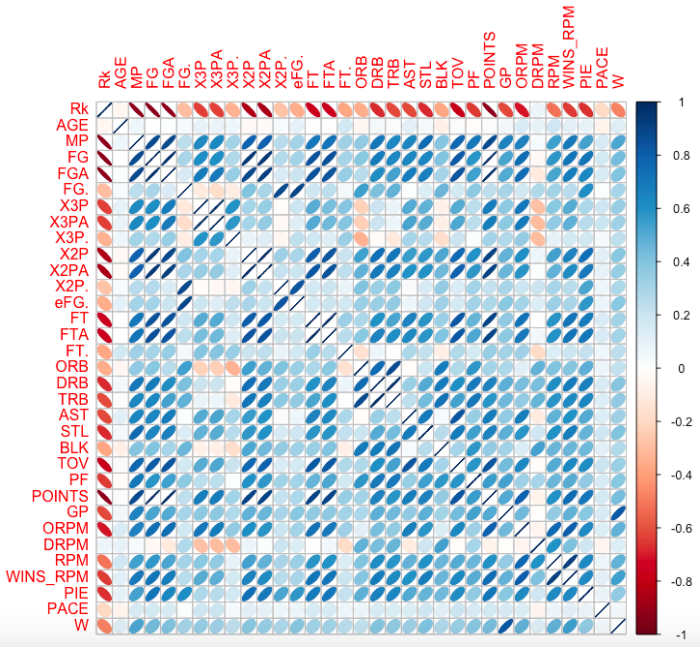
Position	C1	C2	C3
C	5	53	9
PF	7	28	35
PF-C	0	0	1
PG	23	0	47
SF	14	7	42
SG	15	1	58

CA-5C: Top 20 NBA Players



Appendix 6 (Ridge Regression)

RR-1: Correlation Plot



RR-3A: Linear Regression Result

```
Call:
lm(formula = WINS_RPM ~ AGE + MP + X3P + X2PA + eFG. + ORB +
    DRB + AST + STL + BLK + TOV + PF + POINTS + GP + ORPM + DRPM +
    PIE + PACE + W, data = nba)

Residuals:
   Min     1Q Median     3Q    Max
-2.767 -0.529 -0.065  0.516  3.686

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  6.11824    1.55268    3.94      0.0000950284563 ***
AGE         -0.02127    0.01039   -2.05              0.04127 *
MP          -0.07398    0.02031   -3.64              0.00030 ***
X3P         -0.90429    0.19866   -4.55      0.0000069444863 ***
X2PA        -0.51401    0.07473   -6.88      0.0000000000217 ***
eFG.         1.44899    0.87176    1.66              0.09722 .
ORB          0.41723    0.12451    3.35              0.00088 ***
DRB          0.44820    0.06476    6.92      0.0000000000165 ***
AST          0.63658    0.07300    8.72 < 0.0000000000000002 ***
STL          0.74472    0.20123    3.70              0.00024 ***
BLK          0.39690    0.17538    2.26              0.02413 *
TOV         -0.84893    0.18085   -4.69      0.0000036173293 ***
PF          -0.82263    0.11330   -7.26      0.0000000000018 ***
POINTS       0.65016    0.06310   10.30 < 0.0000000000000002 ***
GP           0.01201    0.00422    2.85              0.00463 **
ORPM         0.79755    0.05761   13.84 < 0.0000000000000002 ***
DRPM         0.97636    0.04534   21.53 < 0.0000000000000002 ***
PIE         -0.28990    0.03345   -8.67 < 0.0000000000000002 ***
PACE        -0.03601    0.01544   -2.33              0.02018 *
W            0.01306    0.00572    2.28              0.02281 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.89 on 426 degrees of freedom
Multiple R-squared:  0.943,     Adjusted R-squared:  0.941
F-statistic:  374 on 19 and 426 DF,  p-value: <0.0000000000000002
```

```
Call:
lm(formula = WINS_RPM ~ ., data = nba)

Residuals:
   Min     1Q Median     3Q    Max
-4.162 -0.533 -0.105  0.497  3.440

Coefficients:
              Estimate    Std. Error t value             Pr(>|t|)
(Intercept)  3.9335932905  1.6078666201    2.45              0.01483 *
salary       0.0000000216  0.0000000102    2.13              0.03402 *
AGE         -0.0306791806  0.0115533077   -2.66              0.00822 **
X3P          0.4749900993  0.1019323760    4.66      0.0000042405126229 ***
X3P.         0.5096367945  0.4257457434    1.20              0.23196
X2P.         0.8592686101  0.6379249567    1.35              0.17871
FTA          0.3976693075  0.0491763915    8.09 < 0.0000000000000065 ***
FT.          0.1248168219  0.3218103441    0.39              0.69832
ORB          0.5910007489  0.1146148146    5.16      0.0000003866575968 ***
AST          0.2656516825  0.0450120440    5.90      0.0000000073564151 ***
STL          0.4928289848  0.1872745902    2.63              0.00881 **
BLK          0.6235864375  0.1766833476    3.53              0.00046 ***
PF          -0.7485359977  0.1133856604   -6.60      0.0000000001214752 ***
GP           0.0117710160  0.0041925895    2.81              0.00522 **
RPM          1.0099274805  0.0323010259   31.27 < 0.0000000000000002 ***
PIE         -0.1711994137  0.0260834289   -6.56      0.0000000001533340 ***
PACE        -0.0146049688  0.0160042373   -0.91              0.36199
W            0.0115788056  0.0059955275    1.93              0.05412 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.94 on 425 degrees of freedom
Multiple R-squared:  0.928,     Adjusted R-squared:  0.925
F-statistic:  323 on 17 and 425 DF,  p-value: <0.0000000000000002
```

(1) - Linear regression result left backward right manual - WINS_RPM

```
Call:
lm(formula = POINTS ~ FG + X3P + X3PA + X2P + X2PA + eFG. + FT +
    DRPM + WINS_RPM + PIE, data = nba)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19527 -0.06545  0.00484  0.04552  0.20198

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) -0.08144    0.02943   -2.77              0.00589 **
FG           1.69509    0.07503   22.59 < 0.0000000000000002 ***
X3P          1.19114    0.07821   15.23 < 0.0000000000000002 ***
X3PA         0.03563    0.01265    2.82              0.00509 **
X2P          0.26177    0.07478    3.50              0.00051 ***
X2PA         0.02535    0.00793    3.20              0.00149 **
eFG.         0.18319    0.06770    2.71              0.00708 **
FT           0.98571    0.00536  184.01 < 0.0000000000000002 ***
DRPM        -0.00750    0.00353   -2.12              0.03426 *
WINS_RPM     0.00858    0.00229    3.75              0.00020 ***
PIE         -0.00280    0.00185   -1.51              0.13059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.073 on 435 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:     1
F-statistic: 3.1e+05 on 10 and 435 DF,  p-value: <0.0000000000000002
```

```
Call:
lm(formula = POINTS ~ ., data = nba)

Residuals:
   Min     1Q Median     3Q    Max
-7.887 -0.937 -0.082  0.782  5.023

Coefficients:
               Estimate    Std. Error t value             Pr(>|t|)
(Intercept)  2.6357586583  2.6603100537    0.99              0.32236
salary       0.0000000771  0.0000000165    4.67           0.00000411 ***
AGE         -0.0540886917  0.0191470881   -2.82              0.00495 **
X3P          3.0318032775  0.1663121238   18.23 < 0.0000000000000002 ***
X3P.        -0.2024189878  0.7036654667   -0.29              0.77374
X2P.        -0.2079481362  1.0520915486   -0.20              0.84341
FTA          1.7280121566  0.0812971431   21.26 < 0.0000000000000002 ***
FT.          0.6640093017  0.5328162104    1.25              0.21337
ORB          0.7005044791  0.1890704696    3.70              0.00024 ***
AST          0.1812313255  0.0737134228    2.46              0.01435 *
STL          0.6357824847  0.3089895166    2.06              0.04024 *
BLK          0.4602744340  0.2906842761    1.58              0.11407
PF           0.3069196459  0.1866254085    1.64              0.10080
GP           0.0068219726  0.0069186835    0.99              0.32468
RPM         -0.2325892877  0.0529362371   -4.39           0.00001409 ***
PIE          0.2257238149  0.0429446972    5.26           0.00000023 ***
PACE        -0.0383373473  0.0264483348   -1.45              0.14793
W            0.0035227541  0.0098952969    0.36              0.72202
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.6 on 425 degrees of freedom
Multiple R-squared:  0.937,     Adjusted R-squared:  0.935
F-statistic:  373 on 17 and 425 DF,  p-value: <0.0000000000000002
```

(2) - Linear regression result left backward right manual - Points

RR-3B: RMSE, R-Squared and MAE Table

```
> lm$results
  intercept      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
1      TRUE 0.8899363 0.9373185 0.6932408 0.1286743 0.02249583 0.09925221
```

(1)- Linear regression results _WINS_RPM _Backward

```
> lm$results
  intercept      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
1      TRUE 0.9985843 0.9167636 0.7561949 0.1553921 0.03633877 0.1068711
```

(2) - Linear regression results _WINS_RPM _Manual

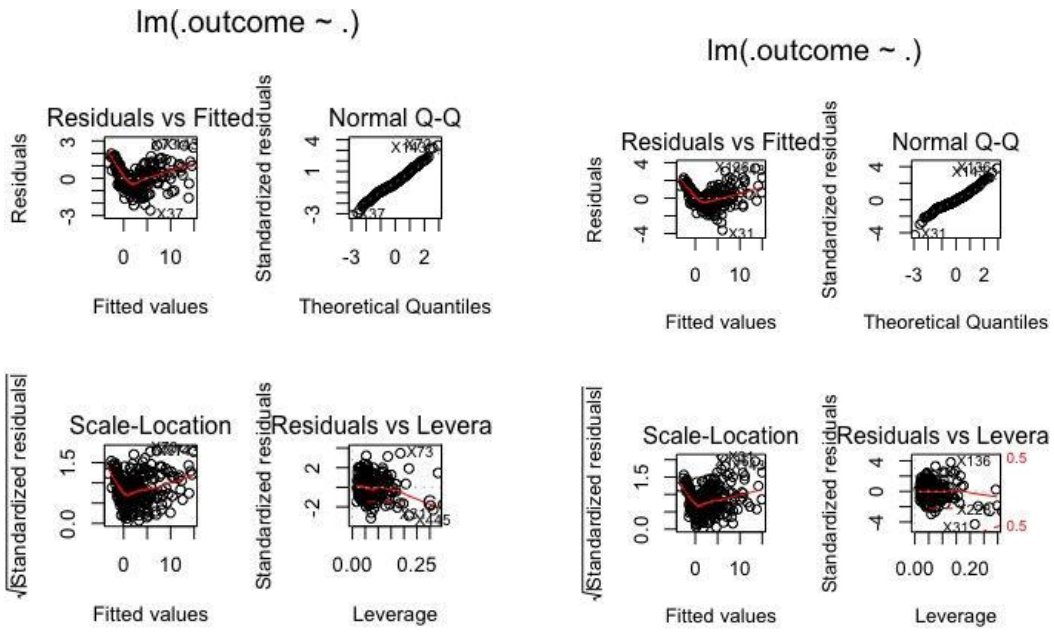
```
> lm$results
  intercept       RMSE  Rsquared        MAE      RMSESD   RsquaredSD       MAESD
1      TRUE 0.07153011 0.9998562 0.05800879 0.006390458 5.145949e-05 0.006610723
```

(3) - Linear regression results_Points_Backward

```
> lm$results
  intercept      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
1      TRUE 1.619754 0.9321526 1.185346 0.2915828 0.02140585 0.1669579
```

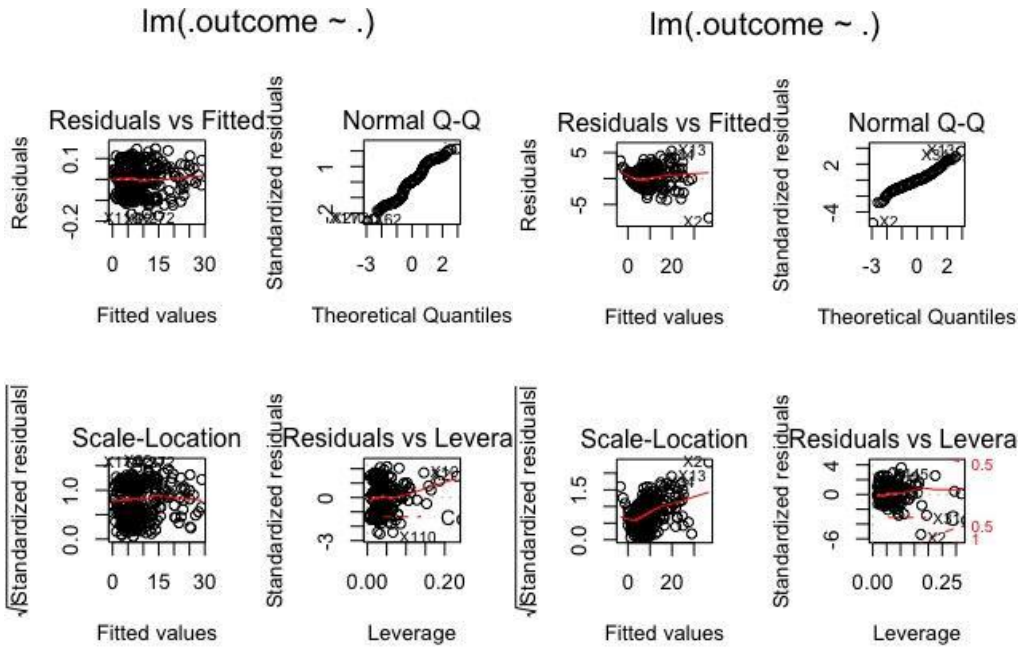
(4) - Linear regression results_Points_Manual

RR-2: Residuals Plots



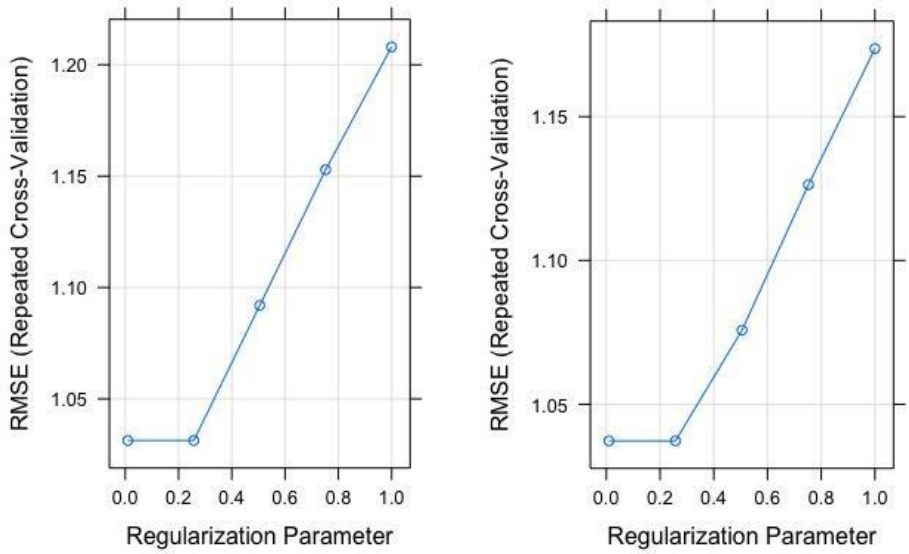
(1)- Residual plots left backward right manual-WINS_RPM

RR-4: Residual Plots

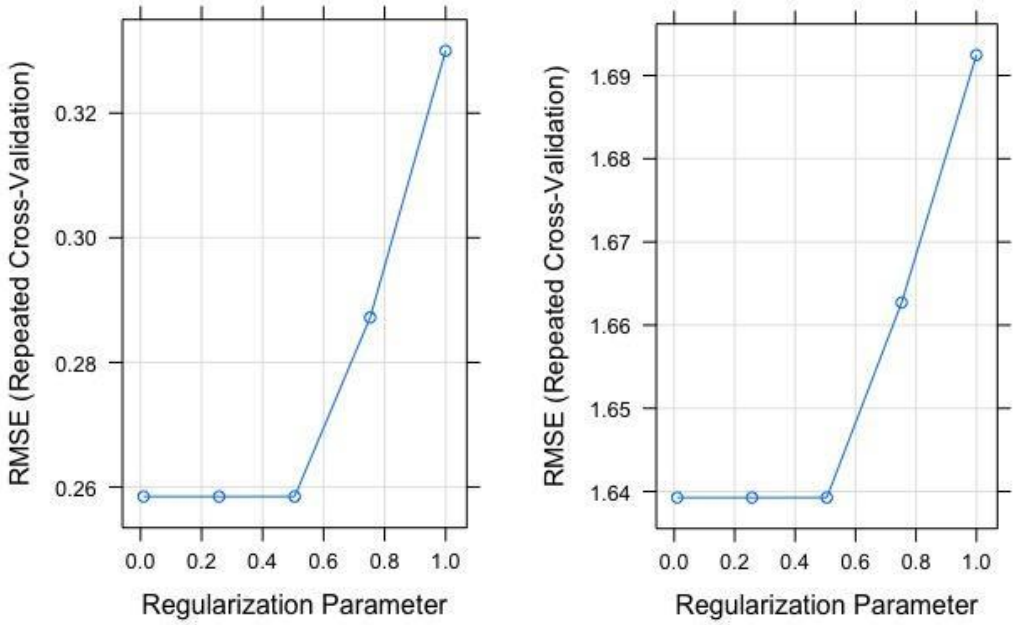


(2)- Residual plots left backward right manual- Points

RR-5: Ridge Regression Plot

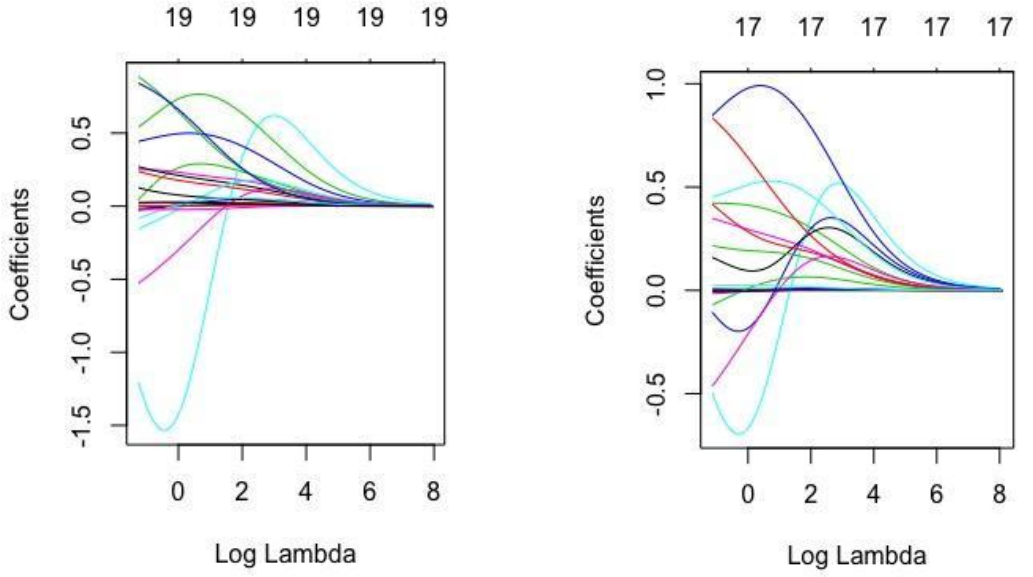


(1) Ridge Regression-RMSE v.s Regularization plot left backward right manual- WINS_RPM

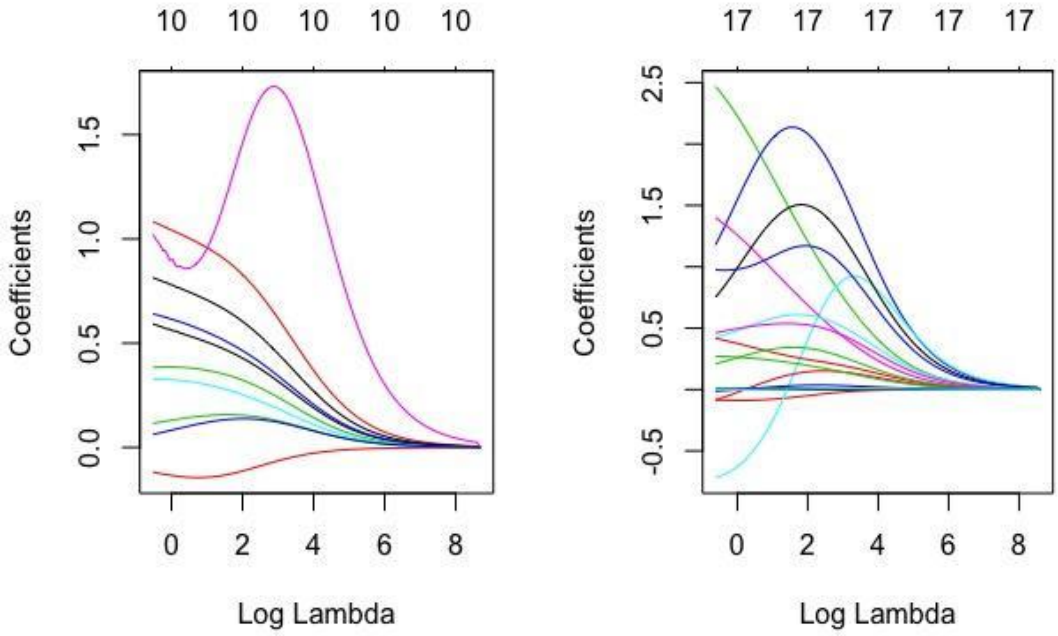


(2) Ridge Regression-RMSE v.s Regularization plot left backward right manual - Points

RR-6: Log Lambda Plot

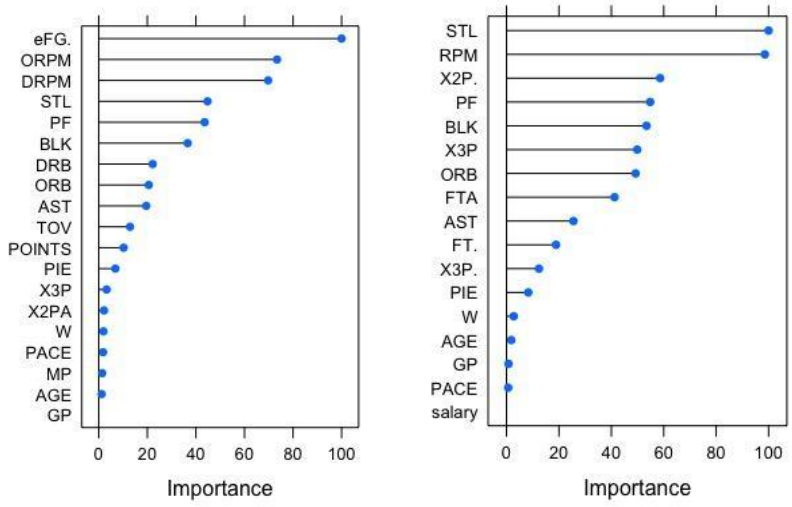


(1) Ridge Regression -log lambda plot left backward right manual - WINS_RPM

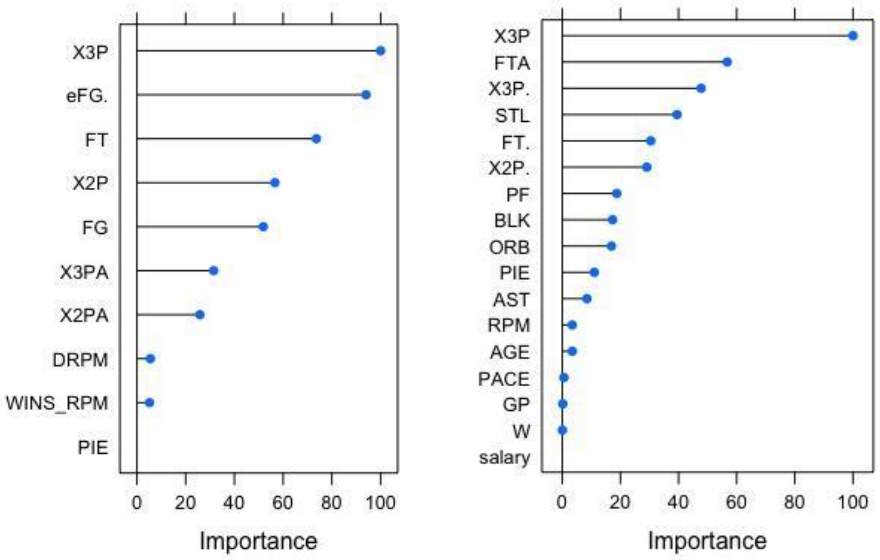


(2) Ridge Regression -log lambda plot left backward right manual - Points

RR-7A: Important Variables for Ridge Regression Figure



(1) - Important variables- Ridge(standardized) left backward right manual -WINS_RPM



(1) - Important variables- Ridge(standardized) left backward right manual- Points

RR-7B: Important Variables for Ridge Regression Table

```
MAE
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
linearmodel 0.4957068 0.6102804 0.7045035 0.6932408 0.7547726 0.9951591    0
Ridge       0.5113883 0.6937910 0.7601818 0.7725828 0.8283131 1.1784410    0

RMSE
                Min.   1st Qu.    Median      Mean   3rd Qu.     Max. NA's
linearmodel 0.5907293 0.7996783 0.8868327 0.8899363 0.9681785 1.336377    0
Ridge       0.6831556 0.8902880 0.9829360 1.0312981 1.1633805 1.748210    0

Rsquared
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
linearmodel 0.8685871 0.9250318 0.9462177 0.9373185 0.9524699 0.9706440    0
Ridge       0.8239152 0.9006208 0.9254703 0.9182291 0.9373138 0.9576222    0
```

(1) Model Comparison- Linear & Ridge -WINS_RPM backward

```
MAE
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
linearModel 0.5162722 0.6973449 0.7540385 0.7561949 0.8288498 0.9710357    0
Ridge       0.5146494 0.6844625 0.7749235 0.7663425 0.8353104 1.0211202    0

RMSE
                Min.   1st Qu.    Median      Mean  3rd Qu.     Max. NA's
linearModel 0.6908989 0.8885071 0.9909082 0.9985843 1.108492 1.369502    0
Ridge       0.6356327 0.9362909 1.0410556 1.0372659 1.155055 1.458980    0

Rsquared
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
linearModel 0.8218254 0.8968883 0.9255980 0.9167636 0.9435055 0.9665178    0
Ridge       0.7928049 0.8919534 0.9244594 0.9102459 0.9444953 0.9637742    0
```

(2) Model Comparison- Linear & Ridge -WINS_RPM manual

```
MAE
                Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
linearmodel 0.04470467 0.05348326 0.05731178 0.05800879 0.06266827 0.07275433    0
Ridge       0.12922324 0.17353599 0.18888567 0.19396190 0.21596358 0.27289328    0

RMSE
                Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
linearmodel 0.05997252 0.06676647 0.07083042 0.07153011 0.07664425 0.08638793    0
Ridge       0.16918381 0.23098266 0.25250253 0.25847965 0.28587822 0.37097757    0

Rsquared
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
linearmodel 0.9997182 0.9998298 0.9998658 0.9998562 0.9998912 0.9999366    0
Ridge       0.9968525 0.9981623 0.9986174 0.9985234 0.9990656 0.9994847    0
```

(3) Model Comparison- Linear & Ridge -Points backward

```
MAE
                 Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
linearmodel 0.7800206 1.079331 1.179104 1.185346 1.280970 1.554582    0
Ridge       0.8831939 1.107431 1.242985 1.222048 1.324263 1.566240    0

RMSE
                Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
linearmodel 1.000923 1.449163 1.620316 1.619754 1.717395 2.532027    0
Ridge       1.073470 1.466927 1.640033 1.639272 1.818171 2.438867    0

Rsquared
                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
linearmodel 0.8808686 0.9202214 0.9290300 0.9321526 0.9467018 0.9855881    0
Ridge       0.8862732 0.9192942 0.9288423 0.9300838 0.9457913 0.9808109    0
```

(4) Model Comparison- Linear & Ridge - Points Manual

RR-8: Model Comparison Figure

```
> ##best model
> ridge$bestTune
  alpha lambda
2     0 0.2575
```

Figure 9 (1) Best Model- Ridge Regression - WINS_RPM

```
> ##best model
> ridge$bestTune
  alpha lambda
3     0  0.505
```

(2) Best Model- Ridge Regression - Points

RR-9: Best Model Ridge Regression Table

	RMSE	R Squared	MAE
Backward	0.89	0.93	0.69
Manual	0.998	0.92	0.756

Table 1 - Linear regression result - WINS_RPM

Backward	eFG.	STL	PF	BLK	DRM	ORM
Manual	STL	RPM	X2P	PF	BLK	ORM

Table 2 - Variables Importance - WINS_RPM

Appendix RR-10: Regression Prediction for "POINTS":

The multiple linear regression results are shown in Appendix 5 Figure 2 &3 - Linear regression results. Table 3 presents the RMSE, R-Squared and MAE values for both using backward selection and manual selection. The RMSE(root mean square error) for POINTS is 0.072 and r squared is 1 which indicates that almost 100% of variability seen in response (POINTS) is explained by backward selection model. The MAE(mean absolute error) is 0.058 for backward selection. The RMSE(root mean square error) for POINTS is 1.61 and R-squared is 0.93 which indicates that almost 93% of variability seen in response (POINTS) is explained by the manual selection model. The MAE is 1.18 for the manual model.

The residual plots are shown in Appendix 5 Figure 4- Residual Plots. Appendix 5 Figure 5- Ridge regression-plot shows the results of ridge regression. On Y-axis the root mean square error which has been estimated using repeated cross-validation. On X-axis we have lambda and RMSE increases as the lambda increases. The minimum RMSE generated when the lambda value is 0.505. The Appendix 6 Figure-6 Ridge Regression -log lambda plot shows the log lambda value on X-axis and coefficients on Y-axis. So when log lambda is about 8, all the coefficients are more or less zero and as lambda is released coefficients starts to grow. As coefficients start to grow, the sum of the square of coefficients becoming larger. The figures show that we have all 10 variables on each point for backward selection model and 17 for manual selection model. Appendix 5 Figure 7- Important variables-Ridge(standardized) shows the plot of variable importance. Table 4 shows that the most important variables which impact the points scored per game are X3P, eFG., FT, X2P, FG, and X3PA for backward selection model; X3P, FTA, X3P%, STL, FT% and X2P% for manual selection model.

The equations for linear regression:

*Points (Backward) = -0.08 +1.70*FG + 1.19 X3P +0.03*X3PA + 0.26*X2P + 0.03*X2PA + 0.18*eFG. +0.98*FT*

*Points (Manual) = 2.64 - 0.054*Age + 3.03 *X3P – 0.2*X3% - 0.21*X2% +1.73*FTA + 0.66*FT% + 0.70*ORB + 0.18*AST + 0.64 STL + 0.46*BLK + 0.31*PF – O.23RPM + 0.23*PIE - 0.04*PACE*

The variables with underlines are shared variables from both backward selection and manual selection which have impact on the points scored per game (POINTS).

Appendix 5 Figure 8 Model Comparison- Linear & Ridge shows the comparison between the linear regression and ridge regression results. It gives a summary in the form of min, 1st quartile, median,

mean, 3rd quartile and maximum. The mean values of RMSE for both the models differentiate significantly. As the differences are very large from both models for backward selection, it is easy to pick Ridge Regression model over the linear regression model for backward selection. As the differences are very small from both models from manual selection, it is difficult to find Ridge Regression model is superior to the linear regression model. Appendix 5 Figure 9 Best Model- Ridge Regression shows that the best model of the ridge is at alpha is zero and lambda is 0.505.

	RMSE	R Squared	MAE
Backward	0.072	1	0.058
Manual	1.61	0.93	1.18

Table 3 - Linear regression result - Points

Backward	X3P	eFG.	FT	X2P	FG	X3PA
Manual	X3P	FTA	X3P%	STL	FT%	X2P%

Table 4 - Variables Importance - Points