DePaul University – Summer, 2020
DSC 423 Class Project - Milestone 10:

Final Report

Group Name: Team Ysgard
Group Members:
Kyle Kassen, Serena Yang,
Agraj Allola, Pramathesh Shukla

Table of Contents

Introduction **3**

Data Preparation **3**

Data Analysis **6**

Module Building **10**

Introduction 10

4.1 – Kyle Kassen 11

4.2 – Serena Yang 20

4.3 – Agraj Allola 43

4.4 – Pramathesh Shukla 58

Discussion **65**

Conclusions **65**

Appendix **66**

Introduction

The National Basketball Association (NBA) is a legendary sports league composed of a variety of players. The league has a history of almost 100 years, and it is challenging to determine a player's value if he were to compare with a player from a different year. Every player has an alternate range of abilities and is necessary to shape a decent group and dominate matches. Each player has a different skill set and is essential to form a good team and win games. The issue arises when there is a need for evaluation for all the different players. For this project, we will use the dataset from the five-thirty-eight website to 1) gain an objective insight into NBA player 'value' and 2) explore if that 'value' impacts team success.

Data Preparation

During our first weekly meeting, we explored several datasets on Kaggle.com and FiveThirtyEight.com that aligned with our assigned topic — sports/marketing. We downloaded and reviewed several datasets with the intention of locating a dataset that had many features, lots of rows, and a continuous response variable. After some discussion and a unanimous group decision, we chose the historical_RAPTOR_by_team dataset from FiveThirtyEight.com. The initial dataset included 27,371 observations and a majority of the variables were quantitative and continuous. Most importantly, we chose a quantitative and continuous response variable; war_total.

Our initial dataset historical_RAPTOR_by_team was clean, so we did not need to do any data preparation. That said, our group decided to seek out a second dataset [Composite ELO] to add more variables in order to fully realize our goal of exploring an objective insight into each player's value. The second dataset was embedded inside of a webpage on FiveThrityEight.com that had to be extracted to a text file before transfer to a CSV file. The data was then manually merged with the original dataset; a long and tedious process. In the end, we added several new variables during the merge process, including; four quantitative independent variables that could be important in explaining or predicting the 'effect' of our dependent variable. During the merge process, we lost about 3,000 observations because of 'season' inconsistencies between the two datasets. The final dataset [RAPTOR_CLEAN] included 24,203 observations and 23 variables.

A full listing of the variables and important methodology equations within the dataset are listed below.

1.1 The Qualitative (Categorical) Variables:

TYPE	Variable	Description
Categorical	**player_name**	Player name
Categorical	**player_id**	Basketball-Reference.com player ID
Categorical, Interval	**season**	Season
Categorical,Binary	**season_type**	Regular season (RS) or playoff (PO)
Categorical	**team**	Basketball-Reference ID of team
Categorical	**Team**	Full Team Name

1.2 The Quantitative Variables:

Independent Variables (Explanatory Variables, Predictive Variables): These are the independent variables we will consider for the project. We assume these variables will help explain/predict potential trends for the response variable war_total.

TYPE	Variable	Description
Continuous	**raptor_offense**	Points above average per 100 possessions added by player on offense, using both box and on-off components
Continuous	**raptor_defense**	Points above average per 100 possessions added by player on defense, using both box and on-off components
Continuous	**raptor_total**	Points above average per 100 possessions added by player on both offense and defense, using both box and on-off components
Continuous	**war_reg_season**	Wins Above Replacement for regular season
Continuous	**war_playoffs**	Wins Above Replacement for playoffs
Continuous	**poss**	Possessions played
Continuous	**mp**	Minutes played
Continuous	**pace_impact**	Player impact on team possessions per 48 minutes
Numeric	**Team Rank**	Ranking Based on ELO score
Numeric	**Peak**	Peak ELO Rating
Numeric	**Mean**	Mean ELO Rating
Rank	**Compsite ELO**	An average of Mean,Peak, and season ending ELO

1.3 - Measures included in the raptor_offense, raptor_defense, and raptor_total variables:

TYPE	Variable	Description
Continuous	**raptor_box_offense**	Points above average per 100 possessions added by player on offense, based only on box score estimate
Continuous	**raptor_box_defense**	Points above average per 100 possessions added by player on defense, based only on box score estimate
Continuous	**raptor_box_total**	Points above average per 100 possessions added by player, based only on box score estimate
Continuous	**raptor_onoff_offense**	Points above average per 100 possessions added by player on offense, based only on plus-minus data
Continuous	**raptor_onoff_defense**	Points above average per 100 possessions added by player on defense, based only on plus-minus data
Continuous	**raptor_onoff_total**	Points above average per 100 possessions added by player, based only on plus-minus data

1.4 Measures of Prediction:

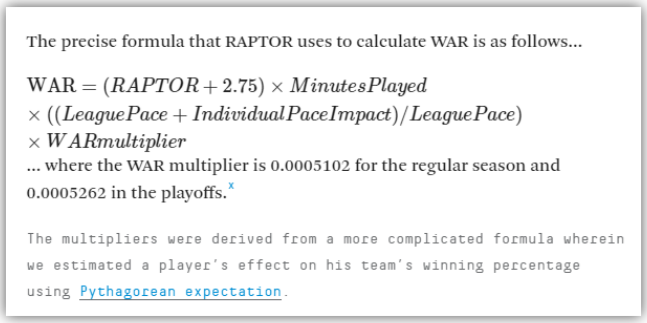
TYPE	Variable	Description
Continuous	**predator_offense**	Predictive points above average per 100 possessions added by player on offense
Continuous	**predator_defense**	Predictive points above average per 100 possessions added by player on defense
Continuous	**predator_total**	Predictive points above average per 100 possessions added by player on both offense and defense

2.1 RAPTOR— Robust **A**lgorithm (using) **P**layer **T**racking (and) **O**n/Off **R**atings**:** This is the general methodology used by Nate Silver and FiveThirtyEight.com. It uses a mixture of traditional statistics, popular sports metrics, and newer methods developed at FiveThirtyEight.com. It's indicative of the variables within the dataset.

Dependent Variable (Response Variable): We have chosen **war_total** as the Dependent Variable (Response Variable). War_total is a Quantitative and Continuous. The variable is a measurable-calculation of other Quantitative variables within the dataset which make it a solid candidate as the response variable.

[war_tota measures 'Wins Above Replacement' (WAR)] — WAR measures a player's value in all facets of the game by deciphering how many more wins he's worth than a replacement-level player at his same position[1].

Figure 1.0 —WAR Formula


The precise formula that RAPTOR uses to calculate WAR is as follows...

$$\text{WAR} = (RAPTOR + 2.75) \times MinutesPlayed \times ((LeaguePace + IndividualPaceImpact)/LeaguePace) \times WARmultiplier$$

... where the WAR multiplier is 0.0005102 for the regular season and 0.0005262 in the playoffs.[x]

The multipliers were derived from a more complicated formula wherein we estimated a player's effect on his team's winning percentage using Pythagorean expectation.

Figure 1.1—Pythagorean Expectation

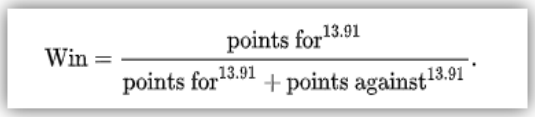
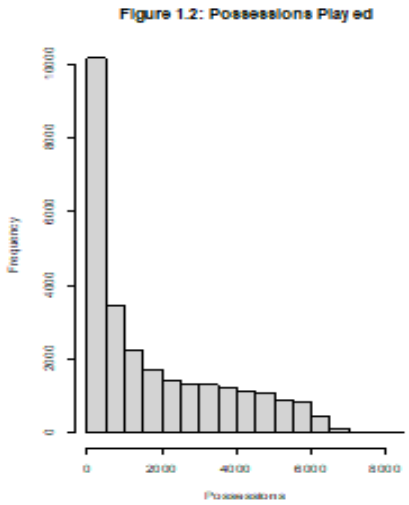

$$\text{Win} = \frac{\text{points for}^{13.91}}{\text{points for}^{13.91} + \text{points against}^{13.91}}.$$

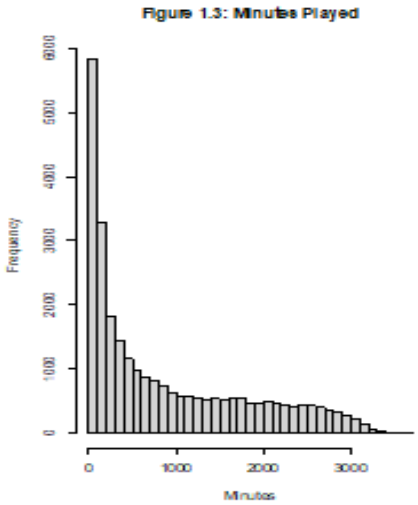
Figure 3.1—ELO Rating System

ELO	EQUIVALENT RECORD	TEAM DESCRIPTION
1800	67-15	All-time great
1700	60-22	Title contender
1600	51-31	Playoff bound
1500	**41-41**	**Average**
1400	31-51	In the lottery
1300	22-60	LOL
1200	15-67	Historically awful

Data Analysis

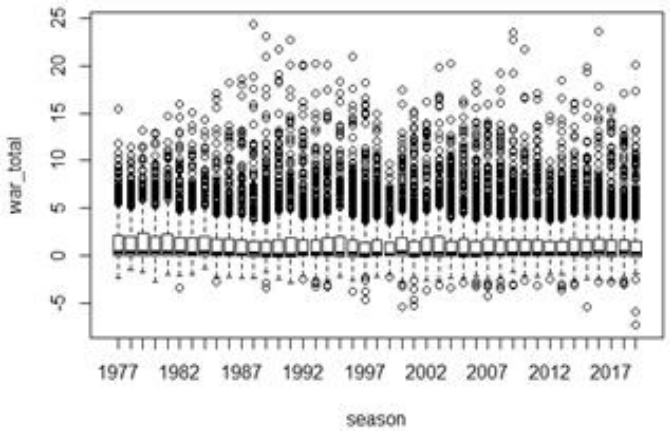
One of the season numbers shows in the dataset, we can consider that as time goes on, years show up increasingly more often. We accept that in this procedure of basketball history, the number of players has expanded as more teams joined the league. In figure 1.2, a large portion of the information is stacked on the left half of the x-axis, which implies that half of the players at the most essential level, a fourth of the players have played somewhere in the range of 400 and 2000 possessions for each season. Only a few players in history have played more than 5000 possessions on the court each year. We will utilize different regression models that incorporate possessions played and minutes played as explanatory variables to show the average time per possession of every player on the court. Notwithstanding, from our histogram analysis of minutes played result in figure 1.3, we can conclude that the histogram's general pattern resembles the past figure. We were surprised to find that the histograms of average per 100 possessions included by the player offense, protection, and both offense and guard are very comparative. At the point when we take a gander at the appropriation in more detail, there are just a handful of good players, which make perfect sense.



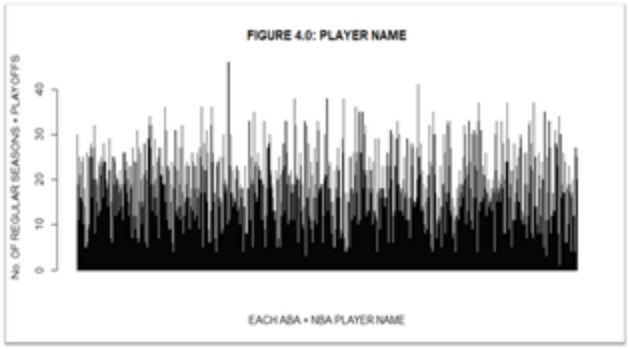


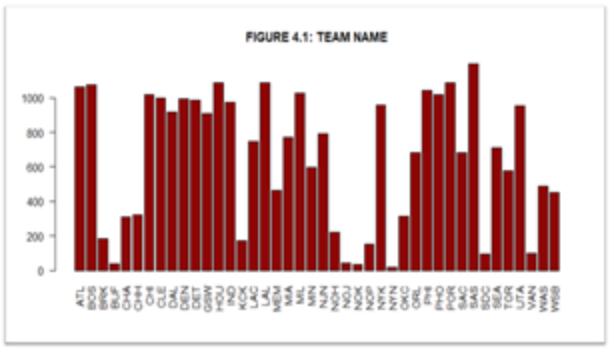
After removing all the categorical and ordinal variables in the five-number summary procedure, the resulting figure 2.0 is apparent. Since when we merge the original file with another dataset and the dataset is not intact from 2016 to 2019, there are 3168 NA's in the last three column variables. From possessions per season and minutes played, since the mean is on the left of the distribution, we can reason that the distribution slanted left and most players are at the basic level. Each of the rest variables' maximum worth could assist us in determining the best players. From this result, all the numbers appear to be reasonable, and no outlier found.

A box plot is created from five characteristics: the minimum value, the primary quartile, the median, the third quartile, and the most extreme worth. We use these characteristics to consider how close other data regards are to them. To build a boxplot, use a flat or vertical number line and a rectangular box. The smallest and biggest data regards mark the endpoints of the axis. The first quartile marks one completion of the box and the third quartile indicates the contrary completion of the crate. Generally, the middle 50 percent of the data falls inside the box. The "whiskers" connect from the completion of the case to the smallest and biggest data esteems. The middle or second quartile can be between the first and third quartiles, or it might be one, or the other, or both. The boxplot gives a nice, quick picture of the data. Somewhere in the range of 1987 and 1992 and somewhere in the range of 2007 and 2012 the war_total was higher

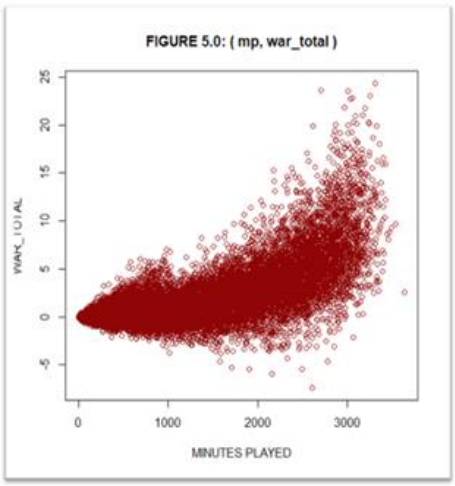


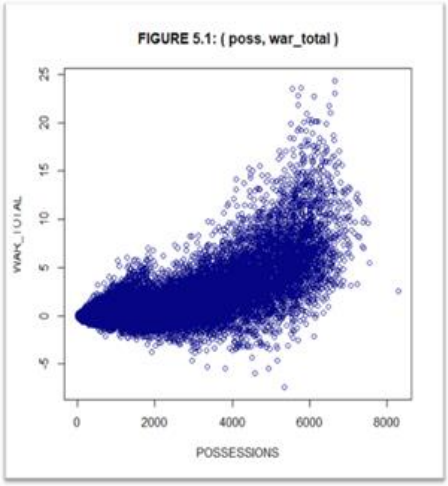
Upon visual review of the bar graph in figure 4.0, we see 'every ABA and NBA player name' on the x-axis and 'number of regular seasons + number of playoffs' on the y-axis. This is a significant categorical variable since we are attempting to explore a target proportion of NBA player 'value'. There is an outlier in the left-central portion of the bar chart; one player has played 40+ regular seasons + playoffs. Upon visual inspection of the bar graph in figure 4. 1, we see 'every ABA and NBA team abbreviation' on the x-axis and 'number of regular seasons + number of playoffs per player' on the y-axis. We watch some huge variations in the numbers on the y-axis since certain groups have not kept up their franchise over time. We likewise notice that many of the more notable urban communities with bigger populations report bigger numbers on the y-axis.





Upon visual inspection of the scatter plot in figure 5. 0, we see a strong positive linear relationship between minutes played and war total. Also, in the scatter plot figure 5. 1, we see a strong positive linear relationship between possessions and war total. The scatter plots in figures 5. Upon visual inspection of the scatter plot in figure 5. 2, we see a moderate positive linear relationship between composite Elo and war total. From the scatter plot in figure 5. Based on the scatterplots in Figure 5. Based on the scatterplots in Figure 5.





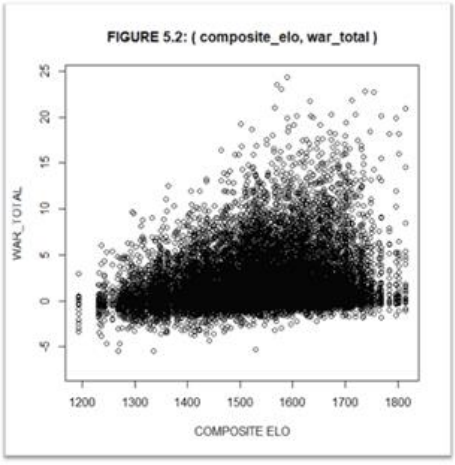
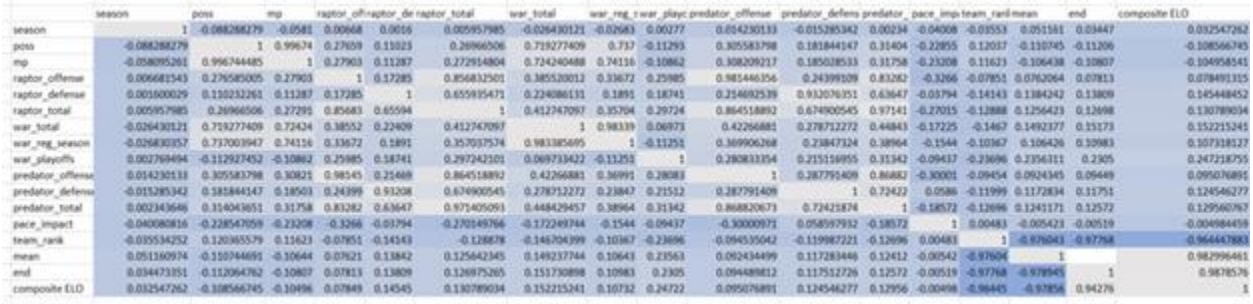


Figure 6. 1 shows the correlation matrix between the various variables of the dataset. We have used ordinal, numeric, and integer data type variables to correlate with each other. For categorical variables, we have used a pivot table. The range of the correlation matrix is. As we can observe the correlation of the same set of variables is 1 as they have the same values and the correlation between them is 1. The correlation of the set of variables has a positive impact if they are 1 or close to 1, similarly if the set of variables have a value -1 or close to -1, they have a negative impact on the evaluation. The correlation of composite ELO and mean is ~0. 98 which is almost equal to 1 so we can conclude that the correlation of that set of variables is very good in the model.



	season	poss	mp	raptor_off	raptor_de	raptor_total	war_total	war_reg_s	war_playc	predator_offense	predator_defens	predator_	pace_impa	team_ranl	mean	end	composite ELO
season	1	-0.088288279	-0.0581	0.00668	0.0016	0.005957985	-0.026430121	-0.02683	0.00277	0.014230133	-0.015285342	0.00234	-0.04008	-0.03553	0.051581	0.03447	0.032547262
poss	-0.088288279	1	0.99674	0.27059	0.11023	0.26966506	0.719277409	0.737	-0.11293	0.305583798	0.181844147	0.31404	-0.22855	0.12037	-0.110745	-0.11206	-0.108566745
mp	-0.058095261	0.996744485	1	0.27903	0.11287	0.272914804	0.724240488	0.74116	-0.10862	0.308209217	0.185028533	0.31758	-0.23208	0.11623	-0.106438	-0.10807	-0.104958541
raptor_offense	0.006681543	0.276585005	0.27903	1	0.17285	0.856832501	0.385520012	0.33672	0.25985	0.985446356	0.24399109	0.83282	-0.3266	-0.07851	0.0762064	0.07813	0.078491315
raptor_defense	0.001600029	0.110232261	0.11287	0.17285	1	0.655935471	0.224086131	0.1891	0.18741	0.254692539	0.932076351	0.63647	-0.03794	-0.14143	0.1384242	0.13809	0.145448452
raptor_total	0.005957985	0.26966506	0.27291	0.85683	0.65594	1	0.412747097	0.35704	0.29724	0.864518892	0.674900545	0.97141	-0.27015	-0.12888	0.1256423	0.12698	0.130789034
war_total	-0.026430121	0.719277409	0.72424	0.38552	0.22409	0.412747097	1	0.98339	0.06973	0.42266885	0.278712272	0.44843	-0.17225	-0.1467	0.1492377	0.15173	0.152215243
war_reg_season	-0.026830357	0.737003947	0.74116	0.33672	0.1891	0.357037574	0.983385695	1	-0.11251	0.369906268	0.23847324	0.38964	-0.1544	-0.10367	0.106426	0.10983	0.107318127
war_playoffs	0.002769494	-0.112927452	-0.10862	0.25985	0.18741	0.297242101	0.069733422	-0.11251	1	0.280833354	0.215516955	0.31342	-0.09437	-0.23686	0.2356311	0.2305	0.247218755
predator_offense	0.014230133	0.305583798	0.30821	0.98545	0.25469	0.864518892	0.42266881	0.36991	0.28083	1	0.287791409	0.86882	-0.30001	-0.09454	0.0924345	0.09449	0.095076891
predator_defense	-0.015285342	0.181844147	0.18503	0.24399	0.91208	0.674900545	0.278712272	0.23847	0.21552	0.287791409	1	0.72422	0.0586	-0.11999	0.1172834	0.11751	0.124546277
predator_total	0.002343646	0.314043651	0.31758	0.83282	0.63647	0.975405091	0.448429457	0.38964	0.31342	0.868820671	0.72421874	1	-0.18572	-0.12696	0.1241171	0.12572	0.129560767
pace_impact	-0.040080816	-0.228547059	-0.23208	-0.3266	-0.01794	-0.270149766	-0.172249744	-0.1544	-0.09437	-0.30000971	0.058597932	-0.18572	1	0.00483	-0.005423	-0.00519	-0.004984459
team_rank	-0.035534252	0.120365579	0.11623	-0.07851	-0.14143	-0.128878	-0.146704399	-0.10367	-0.23696	-0.094535042	-0.119987221	-0.12696	0.00483	1	-0.976043	-0.97768	-0.964447883
mean	0.051360974	-0.110744691	-0.10644	0.07621	0.13842	0.125642345	0.149237744	0.10643	0.23563	0.092434499	0.117283446	0.12412	-0.00542	-0.97604	1		0.982996461
end	0.034473351	-0.112064782	-0.10807	0.07813	0.13809	0.126975265	0.151730898	0.10983	0.2305	0.094489812	0.117512726	0.12572	-0.00519	-0.97768	-0.978945	1	0.9878576
composite ELO	0.032547262	-0.108566745	-0.10496	0.07849	0.14545	0.130789034	0.152215241	0.10732	0.24722	0.095076891	0.124546277	0.12956	-0.00498	-0.96445	-0.97856	0.94276	1

Figure 6.1

Figure 7.1, A pivot table can be used to summarize the number of values in a data variable. We have used it to count the number of players who have played in the Playoff season and Regular season. There are 7882 times players have played in the Playoff season and 19489 in the Regular Season. The number is significant for the regular season. From figure 7. 2, the pivot table displays the number of player records in each team. Here it contains two rows, Team name abbreviation and Count of player_id in the team. We have arranged in alphabetical order for ease of access. We can observe that the San Antonio Spurs team has the highest record of players and the New York Knicks team has the lowest record of players.

Row Labels	Count of player_id
PO	7882
RS	19489
Grand Total	27371

Figure 7.1

Row Label	Count of player_id
ATL	1062
BOS	1076
BRK	186
BUF	37
CHA	309
CHH	322
CHI	1019
CLE	1000
DAL	917
DEN	996
DET	985
GSW	910
HOU	1088
IND	974
KCK	173
LAC	748
LAL	1087
MEM	464
MIA	772
MIL	1026
MIN	598
NJN	790
NOH	222
NOJ	42
NOK	34
NOP	154
NYK	959
NYN	18
OKC	313
ORL	684
PHI	1044
PHO	1017
POR	1087
SAC	682
SAS	1196
SDC	95
SEA	711
TOR	579
UTA	954
VAN	98
WAS	490
WSB	453
Grand Total	**27371**

Figure 7.2

Model Building

Introduction

The overall aim of our project is two-fold: 1) to gain an objective insight into NBA player 'value' and 2) explore if that 'value' has an impact on team success. Using the five-thirty-eight data we hope to answer these questions:

1) Which objective measures determine NBA players 'value'?

2) Do those objective measures of 'value' predict team success?

Team Ysgard has divided the individual efforts by columns; all of which complement one another rather than depend on one another. The team member-model assignments are discussed below in each team-members respective subsection.

Broadly speaking, our methodology included two phases: 1) first-order regression models and 2) second-order regression models.

A Brief Overview of the model-building process:

1) We had to filter and split the master-dataset into two separate datasets — one for regular season the other for the play-offs.
2) We ran the models one-by-one for both regular season and the play-offs.
3) We ran stepAIC for both backward and forward directions for both regular season and play-offs.
4) We validated the 'best' models using n-fold cross-validation for both regular season and play-offs.

The goal for phase 1: *to identify the 'best' first-order model, and then; validate the 'best' first-order model using n-fold cross-validation.*

The goal for Phase 2: *to identify the 'best' second order/interaction model, and then; validate the 'best' second-order/interaction model using n-fold cross-validation.*

Subsection 4.1: Kyle Kassen

4.1.1 first-order regression models

For Phase 1, I will consider the relationship between the response variable — war_total and the explanatory variables raptor_offense and raptor_defense. This will be done separately for the Regular Season and Play-offs. These first-order regression models will be used in relation to goal one with the hope of uncovering valuable insights before attempting to build more complex regression models. raptor_offense and raptor_defense have been identified as two of the potentially most important explanatory variables in predicting our response variable: war_total. We want to further investigate the intricacies of the relationships between these two independent variables and the dependent variable in order to understand their individual contribution in terms of predictive power.

The goal for Phase 1 is to identify the 'best' first-order model, and then; validate the 'best' first-order model using n-fold cross-validation.

FIGURE 1.0

m1	Regular Season	war_total ~ raptor_offense
m2	Regular Season	war_total ~ raptor_defense
m3	Regular Season	war_total ~ raptor_offense + raptor_defense
m4	Play-Offs	war_total ~ raptor_offense
m5	Play-Offs	war_total ~ raptor_defense
m6	Play-Offs	war_total ~ raptor_offense + raptor_defense
m7	Regular Season	war_total ~ mp + poss + raptor_offense + pace_impact
m8	Regular Season	war_total ~ mp + poss + raptor_defense + pace_impact
m9	Regular Season	war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact
m10	Play-Offs	war_total ~ mp + poss + raptor_offense + pace_impact
m11	Play-Offs	war_total ~ mp + poss + raptor_defense + pace_impact
m12	Play-Offs	war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact

The following description represents a concise step-by-step overview of my model building process. The actual experience was not nearly as clean and straightforward. Nevertheless, we learn from this process and keep records so that others may be able to replicate the statistical experience. **The R code and models are located in the appendix under 'appendix 4.1: Kyle Kassen'.** An analysis of the findings is located in the section titled "*Phase 1 Analysis…*".

First, I had to filter and split the master-dataset into two separate datasets — one for Regular Season and the other for Play-Offs. This was done in excel before importing each

separate dataset into R. The remaining steps were done separately for both Regular Season and Play-Offs.

Secondly, I went back to review the scatterplots and correlation matrices. This was done to confirm visual observations of the relationship between the response and explanatory variables which were moderately positive linear relationships for both raptor_offense and raptor_defense. This was then confirmed by reviewing the correlation matrices: [RS] cor(raptor_offense,war_total) = 0.544, cor(raptor_defense,war_total) = 0.363 and [PO] cor(raptor_offense,war_total) = 0.3600114, cor(raptor_defense,war_total) = 0.238484.

Thirdly, I ran the first-order models one-by-one as listed in Figure 1.0 for both Regular Season and Play-Offs. I then reviewed the summary statistics for each model using the summary() function. At this point — I evaluated, compared, contrasted, and made inferences. Ultimately, determining the best model manually.

Next, I confirmed the manual selection process by executing and reviewing stepAIC for both backward and forward directions for both Regular Season and Play-Offs.

Finally, I validated the 'best' model using n-fold cross validation to obtain empirical evidence as to the generalizability and prediction performance of my model selection for both Regular Season and Play-Offs.

The following section summarizes the findings that resulted from the model-building process.

Regular Season:

I began by analyzing models m1, m2, and m3. I noted that all p-values are low and significant at 0.001. All three of these models passed the f-test, (assuming a default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero. I see that m1 has an adj. R^2 of 0.2955 and m2 has an adj. R^2 of 0.1318. Intuitively, I might have expected raptor_offense to be a greater predictor of war_total than raptor_defense because I had reviewed the correlation matrix in the model-building process and recorded that [RS] cor(raptor_offense,war_total) = 0.544, cor(raptor_defense,war_total) = 0.363. I observed the highest adj. R^2 amongst this group of models in m3 at 0.3537 when both raptor_offense and raptor_defense are included in the model. That is, 35.37% of the variability in war_total (y) can be explained by the regression model (m3).

Next, I reviewed models m7, m8, and m9. I did this to observe any changes in the model summary statistics in the presence of other important independent variables. I noted that all p-values are low and significant at 0.001. All three of these models passed the f-test, (assuming a

default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero. Again, I inferred raptor_offense to be a greater predictor of war_total than raptor_defense; noting m7 has an adj. R^2 of 0.5834 and m8 has an adj. R^2 of 0.5676. I observed the highest adj. R^2 amongst this group of models in m9 at 0.606 when both raptor_offense and raptor_defense are included in the model.

Using my knowledge of the dataset I inferred that model m3 is the 'best' first-order model. While I observed that raptor_offense has greater predictive value than raptor_defense, I want to keep both independent variables in the model because including both variables paints a more accurate picture of reality; that is, when determining and measuring the overall 'value' of each player I need to look at both their offensive and defensive contributions and capabilities. Furthermore, while I did see an increase in adj. R^2 in models m7, m8, and m9 — I recognize that keeping these other independent variables in the 'best' first-order model would serve to inflate the adj. R^2 value past the purposes of this individual phase 1 investigation of the independent variables raptor_offense and raptor_defense.

Next, I confirmed the manual selection of model m3 as the 'best' first-order model. First, I ran stepAIC "backward" on m9 to see the stepwise Algorithm's selection results. The results align with our manual selection process with raptor_offense having the highest AIC of 23651 and raptor_defense having the second highest AIC of 23014. Second, I ran stepAIC "forward" on m9 to see the stepwise Algorithm's selection results. Again, the results are aligned with our manual selection process with raptor_offense being added to the model in step 2 and raptor_defense being added to the model in step 3.

Lastly, I used n-fold cross-validation to obtain empirical evidence as to the generalizability and prediction performance of my model selection. As per the acceptable standards, I used ten folds in the cross-validation model. I observed an overall ms of 5.95 which is higher than the residual standard error of 2.434 observed in model m3 which makes me question the predictive performance of our model.

Play-Offs:

I began by analyzing models m4, m5, and m6. I noted that all p-values are low and significant at 0.001. All three of these models passed the f-test, (assuming a default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero. I see that m4 has an adj. R^2 of 0.1295 and m5 has an adj. R^2 of 0.05674. Intuitively, I might have expected raptor_offense to be a greater predictor of war_total than raptor_defense because I had reviewed the correlation matrix in the model-building process and recorded that [PO] cor(raptor_offense,war_total) = 0.3600114, cor(raptor_defense,war_total) = 0.238484. I observed the highest adj. R^2 amongst this group of models in m6 at 0.1576 when both raptor_offense and raptor_defense are included

in the model. That is, 15.76% of the variability in war_total (y) can be explained by the regression model (m6).

Next, I reviewed models m10, m11, and m12. I did this to observe any changes in the model summary statistics in the presence of other important independent variables. I noted that all p-values are low and significant at 0.001 except for pace_impact in m11 which was significant at 0.01. All three of these models passed the f-test, (assuming a default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero. Again, I inferred raptor_offense to be a greater predictor of war_total than raptor_defense; noting m10 has an adj. R^2 of 0.693 and m11 has an adj. R^2 of 0.6813. I observed the highest adj. R^2 amongst this group of models in m12 at 0.7 when both raptor_offense and raptor_defense are included in the model.

Using my knowledge of the dataset I inferred that model m6 is the 'best' first-order model. While I observed that raptor_offense has greater predictive value than raptor_defense, I want to keep both independent variables in the model because including both variables paints a more accurate picture of reality; that is, when determining and measuring the overall 'value' of each player I need to look at both their offensive and defensive contributions and capabilities. Furthermore, while I did see an increase in adj. R^2 in models m10, m11, and m12 — I recognize that keeping these other independent variables in the 'best' first-order model would serve to inflate the adj. R^2 value past the purposes of this individual phase 1 investigation of the independent variables raptor_offense and raptor_defense.

Next, I confirmed the manual selection of model m6 as the 'best' first-order model. First, I ran stepAIC "backward" on m12 to see the stepwise Algorithm's selection results. The results align with our manual selection process with raptor_offense having the highest AIC of -11042 and raptor_defense having the second highest AIC of -11304. Second, I ran stepAIC "forward" on m12 to see the stepwise Algorithm's selection results. Again, the results are aligned with our manual selection process with raptor_offense being added to the model in step 2 and raptor_defense being added to the model in step 3.

Finally, I used n-fold cross-validation to obtain empirical evidence as to the generalizability and prediction performance of my model selection. As per the acceptable standards, I used ten folds in the cross-validation model. I observed an overall ms of 0.551 which is lower than the residual standard error of 0.7411 observed in m6 which gives me confidence in the model predictive performance.

4.2.2 Second-order Regression

For phase 2, I will consider the relationship between the response variable — war_total and the explanatory variables raptor_offense and raptor_defense. This will be done separately for the Regular Season and Play-offs. These regression models will be used in relation to goal one

with the hope of building on the valuable insights gathered in Phase 1 by creating more complex models. raptor_offense and raptor_defense have been identified as two of the potentially most important explanatory variables in predicting our response variable: war_total. We want to further investigate the intricacies of the relationships between these two independent variables and the dependent variable in order to understand their contribution in terms of predictive power.

The goal for Phase 2 is to identify the 'best' second-order/interaction model, and then; validate the 'best' second-order/interaction model using n-fold cross-validation.

FIGURE 2.0

Model 1	Regular Season	`model1 <- lm(war_total ~ raptor_offense + raptor_defense + raptorOD, data=RAPTOR_CLEAN_RS)`
Model 2	Regular Season	`model2 <- lm(war_total ~ raptor_defense + raptor_defense2 + raptor_offense + raptor_offense2,data=RAPTOR_CLEAN_RS)`
Model 3	Regular Season	`model3 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptorOD,data=RAPTOR_CLEAN_RS)`
Model 4	Play-Offs	`model4 <- lm(war_total ~ raptor_offense + raptor_defense + raptorOD, data=RAPTOR_CLEAN_RS)`
Model 5	Play-Offs	`model5 <- lm(war_total ~ raptor_defense + raptor_defense2 + raptor_offense + raptor_offense2,data=RAPTOR_CLEAN_RS)`
Model 5	Play-Offs	`model6 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptorOD,data=RAPTOR_CLEAN_RS)`

The following description represents a concise step-by-step overview of my model building process. The actual experience was not nearly as clean and straightforward. Nevertheless, we learn from this process and keep records so that others may be able to replicate the statistical experience. **The R code and models are located in the appendix 'appendix 4.1: Kyle Kassen'.** An analysis of the findings is located in the section titled "*Phase 2 Analysis*".

First, I had to filter and split the master-dataset into two separate datasets — one for Regular Season and the other for Play-Offs. This was done in excel before importing each separate dataset into R. The remaining steps were done separately for both Regular Season and Play-Offs.

Secondly, I experimented with various second-order terms and interaction terms then I ran the models listed in figure 2.0 for both Regular Season and Play-Offs. I then reviewed the summary statistics for each model using the summary() function. At this point, I evaluated, compared, contrasted, and inferences. Ultimately, determining the best model manually.

Thirdly, I confirmed the manual selection process by executing and reviewing stepAIC for both backward and forward directions for both Regular Season and Play-Offs.

Next, I checked the independent variables for multicollinearity using the cor() function and vif(model) function for both the Regular Season and Play-Offs.

Fifthly, I ran a residual analysis by checking three of the four error assumptions. More specifically: 1) the sum(model$residuals) function to check that the mean is equal to zero, 2) the Durbin-Watson test, and 3) the normal distribution histograms.

Finally, I validated the 'best' model using n-fold cross-validation to obtain empirical evidence as to the generalizability and prediction performance of my model selection for both Regular Season and Play-Offs.

The following section summarizes the findings that resulted from the model-building process.

Regular Season:

I began by analyzing model1, model2, and model3. I noted that all p-values are low and significant at 0.001. All three of these models passed the f-test, (assuming a default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero. I see that model1 had an adj. R^2 of 0.4458 and model2 had an adj. R^2 of 0.4169. For model1 I used an interaction term and for model2 I used a second-order term. Ultimately, I chose to use the interaction term from model1 in model3 because model1 had a higher adj. R^2. Model3 had an adj. R^2 of 0.6434. I made note that all models from M08 had achieved higher adj. R^2 values when compared to the M07 models. Using my knowledge of the dataset I inferred that model1 is the 'best' second-order model; noting, 44.58% of the variability in war_total (y) can be explained by the regression model (model1).

Next, I confirmed the manual selection of model1 as the 'best' second-order model. First, I ran stepAIC "backward" on model3 to see the stepwise Algorithm's selection results. The results align with our manual selection process with our interaction term having an AIC value of 22058. Second, I ran stepAIC "forward" on model3 to see the stepwise Algorithm's selection results. Again, the results aligned with our manual selection process with our interaction term being added to the model in step 4.

Moving along, I used n-fold cross validation to obtain empirical evidence as to the generalizability and prediction performance of my model selection. As per the acceptable standards, I used ten folds in the cross-validation model. I observed an overall ms of 5.15 which is higher than the residual standard error of 1.808 observed in model3 which makes me question the predictive performance of our model.

Finally, I used the cor() function and vif(model) to check for multicollinearity. All independent variables had low vif values well below 10. I finished by running a residual analysis to check three of the four error assumptions. More specifically: 1) the sum(model$residuals) function to check that the mean is equal to zero, 2) the Durbin-Watson test, and 3) the normal distribution histograms. The sum of the errors was low but not exactly

equal to zero; probably a rounding error. The Durbin-Watson test values were satisfactory, having p-values at 0.004 for model1 and 0 for model3. The distributions were normal as per the histograms, but I did note a slight skew to the left for model1 in the non-zscore histogram.

Play-Offs:

I began by analyzing model4, model5, and model6. I noted that all p-values are low and significant at 0.001. All three of these models passed the f-test, (assuming a default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero. I see that model4 had an adj. R^2 of 0.1592 and model5 had an adj. R^2 of 0.1803. For model4 I used an interaction term and for model5 I used a second-order term. Ultimately, I chose to use the second-order term from model5 in model6 because model5 had a higher adj. R^2. Model6 had an adj. R^2 of 0.7045. I made note that all models from M08 had achieved higher adj. R^2 values when compared to the M07 models. Using my knowledge of the dataset I inferred that model5 is the 'best' second-order model; noting, 18.03% of the variability in war_total (y) can be explained by the regression model (model5).

Next, I confirmed the manual selection of model5 as the 'best' second-order model. First, I ran stepAIC "backward" on model6 to see the stepwise Algorithm's selection results. The results align with our manual selection process with our second-order terms having AIC values of -11506 for raptor_defense2 and -11496 for raptor_offense2. Second, I ran stepAIC "forward" on model6 to see the stepwise Algorithm's selection results. Again, the results aligned with our manual selection process with our second-order terms being added to the model in step 3 [rator_offense2] and step 4 [raptor_defense2].

Moving along, I used n-fold cross validation to obtain empirical evidence as to the generalizability and prediction performance of my model selection. As per the acceptable standards, I used ten folds in the cross-validation model. I observed an overall ms of 0.54 which is lower than the residual standard error of 0.731 observed in model5 which makes me confident about the predictive performance of our model.

Finally, I used the cor() function and vif(model) to check for multicollinearity. All independent variables had low vif values well below 10. I finished by running a residual analysis to check three of the four error assumptions. More specifically: 1) the sum(model$residuals) function to check that the mean is equal to zero, 2) the Durbin-Watson test, and 3) the normal distribution histograms. The sum of the errors were low but not exactly equal to zero; probably a rounding error. The Durbin-Watson test values were satisfactory, having p-values at 0 for model5 and 0 for model6. The distributions were normal as per the histograms, but I did note a slight skew to the left for model5 in the non-zscore histogram.

*Best model for **phase 1** - First-order regression:*
Regular Season: war_total ~ raptor_offense + raptor_defense, data = RAPTOR_CLEAN_RS
Play-Offs: war_total ~ raptor_offense + raptor_defense, data = RAPTOR_CLEAN_PO

*Best model for **phase 2** - Second-order regression:*
Regular Season: war_total ~ raptor_offense + raptor_defense + raptorOD, data = RAPTOR_CLEAN_RS
Play-Offs: war_total ~ raptor_defense + raptor_defense^2 + raptor_offense + raptor_offense^2, data = RAPTOR_CLEAN_PO

Both models (m3) and (m6) passed the f-test, (assuming a default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero.
35.37% of the variability in war_total (y) can be explained by the regression model (m3)
15.76% of the variability in war_total (y) can be explained by the regression model (m6)

Both models (model1) and (model5) passed the f-test, (assuming a default alpha of 0.05) so we can reject the null hypothesis that all Betas are equal to zero and accept the alternative that at least one of the Betas is not equal to zero.
44.58% of the variability in war_total (y) can be explained by the regression model (model1)
18.03% of the variability in war_total (y) can be explained by the regression model (model5)

I found that raptor_offense has greater predictive value than raptor_defense for our objective measure of player value. That said, while I observed that raptor_offense has greater predictive value than raptor_defense in determining war_total, I want to keep both independent variables in the model because including both variables paints a more accurate picture of reality; that is, when determining and measuring the overall ‘value’ of each player we need to look at both their offensive and defensive contributions and capabilities. Further, for the regular season in phase 2 an interaction term was chosen for the ‘best’ model while for the play-offs in phase 2 a second-order term was chosen for the ‘best’ model which served to validate our instinct to keep both raptor_offense and raptor_defense in the model. Overall, raptor_offense and raptor_defense were better predictors for war_total in the regular season than in the play-offs for both phases.

Subsection 4.2: Serena Yang

4.2.1 First-order Regression

I will investigate the relationship between the response variable, war_total, and the explanatory variables, raptor_total and pace_impact. This section will be done separately for the Regular Season and Play-offs. These first-order regression models will be used concerning goal one with the hope of uncovering valuable insights before attempting to build more complex regression models.

m12	Regular Season	war_total ~ raptor_total
m13	Regular Season	war_total ~ pace_impact
m14	Regular Season	war_total ~ raptor_total + pace_impact
m15	Play-Offs	war_total ~ raptor_total
m16	Play-Offs	war_total ~ pace_impact
m17	Play-Offs	war_total ~ raptor_total + pace_impact
m18	Regular Season	war_total ~ mp + poss + raptor_total
m19	Regular Season	war_total ~ mp + poss + pace_impact
m20	Regular Season	war_total ~ mp + poss + raptor_total + pace_impact
m21	Play-Offs	war_total ~ mp + poss + raptor_total
m22	Play-Offs	war_total ~ mp + poss + pace_impact
m23	Play-Offs	war_total ~ mp + poss + raptor_total + pace_impact

Regular Season

The first step is by using first-order regression models to find the relationship between the response variable, war_total, and the explanatory variables, raptor_total, and pace_impact for the regular season.

m12 <- lm(war_total ~ raptor_total, data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-18.195  -1.354  -0.818   0.689  36.712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.398871   0.020050  119.65   <2e-16 ***
raptor_total 0.379920   0.003923   96.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.435 on 17171 degrees of freedom
Multiple R-squared:  0.3533,    Adjusted R-squared:  0.3532
F-statistic:  9380 on 1 and 17171 DF,  p-value: < 2.2e-16
```

m13 <- lm(war_total ~ pace_impact, data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.0343 -1.7785 -0.9116  0.8573 22.4599

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.91916    0.02333   82.25   <2e-16 ***
pace_impact -0.88910    0.02468  -36.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.919 on 17171 degrees of freedom
Multiple R-squared:  0.07027,   Adjusted R-squared:  0.07022
F-statistic:  1298 on 1 and 17171 DF,  p-value: < 2.2e-16
```

m14 <- lm(war_total ~ raptor_total + pace_impact, data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-17.094  -1.356  -0.816   0.678  35.868

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.434851   0.020273  120.11   <2e-16 ***
raptor_total  0.364778   0.004164   87.60   <2e-16 ***
pace_impact  -0.231114   0.021849  -10.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.427 on 17170 degrees of freedom
Multiple R-squared:  0.3575,    Adjusted R-squared:  0.3574
F-statistic:  4776 on 2 and 17170 DF,  p-value: < 2.2e-16
```

After deriving the above three results of regression models, we can find that these three F-value are pretty good. By comparing their adjusted R-squared, although the value of m14 is the largest of the three models, 35.74% is still pretty small. At this stage, our team decided by adding two more explanatory variables, mp and poss, to find a better model.

m18 <- lm(war_total ~ mp + poss + raptor_total, data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.6203 -0.9881 -0.1068  0.6578 19.6447

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.901e-01  2.964e-02  -6.414 1.46e-10 ***
mp            2.579e-03  1.629e-04  15.832  < 2e-16 ***
poss         -3.541e-04  8.001e-05  -4.425 9.69e-06 ***
raptor_total  1.893e-01  3.595e-03  52.645  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.913 on 17169 degrees of freedom
Multiple R-squared:  0.6006,    Adjusted R-squared:  0.6006
F-statistic:  8607 on 3 and 17169 DF,  p-value: < 2.2e-16
```

m19 <- lm(war_total ~ mp + poss + pace_impact, data = RAPTOR_CLEAN_RS)

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.2350  -1.0565   0.1237   0.8911  17.6401

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.203e+00  2.890e-02 -41.633  < 2e-16 ***
mp           3.392e-03  1.752e-04  19.363  < 2e-16 ***
poss        -4.973e-04  8.616e-05  -5.773 7.94e-09 ***
pace_impact  9.000e-02  1.895e-02   4.750 2.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.061 on 17169 degrees of freedom
Multiple R-squared:  0.5368,    Adjusted R-squared:  0.5367
F-statistic:  6632 on 3 and 17169 DF,  p-value: < 2.2e-16
```

m20 <- lm(war_total ~ mp + poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.5561 -0.9838 -0.1187  0.6625 19.9277

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.350e-01  3.101e-02 -10.800  < 2e-16 ***
mp            2.706e-03  1.621e-04  16.690  < 2e-16 ***
poss         -3.796e-04  7.952e-05  -4.774 1.82e-06 ***
raptor_total  1.990e-01  3.631e-03  54.798  < 2e-16 ***
pace_impact   2.648e-01  1.777e-02  14.903  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.901 on 17168 degrees of freedom
Multiple R-squared:  0.6057,    Adjusted R-squared:  0.6056
F-statistic:  6594 on 4 and 17168 DF,  p-value: < 2.2e-16
```

After adding two more explanatory variables and comparing three adjusted R-squared, the last regression model is the best. Also, the F-value is pretty good, and each terms' P-values are extremely small. For double check the last regression model for the regular season is the best one at this stage, we ran backward and forward selection processes.

Backward

```
Start:  AIC=22070
war_total ~ poss + mp + raptor_total + pace_impact

              Df Sum of Sq   RSS   AIC
<none>                     62050 22070
- poss         1        82 62132 22091
- pace_impact  1       803 62852 22289
- mp           1      1007 63056 22345
- raptor_total 1     10853 72902 24836
```

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
war_total ~ poss + mp + raptor_total + pace_impact

Final Model:
war_total ~ poss + mp + raptor_total + pace_impact


  Step Df Deviance Resid. Df Resid. Dev   AIC
1                     17168      62050 22070
```

Forward

```
Start:  AIC=38046
war_total ~ 1

               Df Sum of Sq    RSS   AIC
+ mp            1     84245  73134 24887
+ poss          1     82834  74545 25215
+ raptor_total  1     55598 101781 30563
+ pace_impact   1     11060 146319 36796
<none>                      157379 38046

Step:  AIC=24887
war_total ~ mp

               Df Sum of Sq   RSS   AIC
+ raptor_total  1     10210 62924 22307
+ poss          1       135 72998 24857
+ pace_impact   1        90 73044 24868
<none>                    73134 24887

Step:  AIC=22307
war_total ~ mp + raptor_total

              Df Sum of Sq   RSS   AIC
+ pace_impact  1       792 62132 22091
+ poss         1        72 62852 22289
<none>                    62924 22307

Step:  AIC=22091
war_total ~ mp + raptor_total + pace_impact

       Df Sum of Sq   RSS   AIC
+ poss  1      82.4 62050 22070
<none>             62132 22091

Step:  AIC=22070
war_total ~ mp + raptor_total + pace_impact + poss
```

After ran backward and forward selection processes, the final results are the same, which means we are in the right direction. Therefore, for the moment, we set the final model for the regular season as

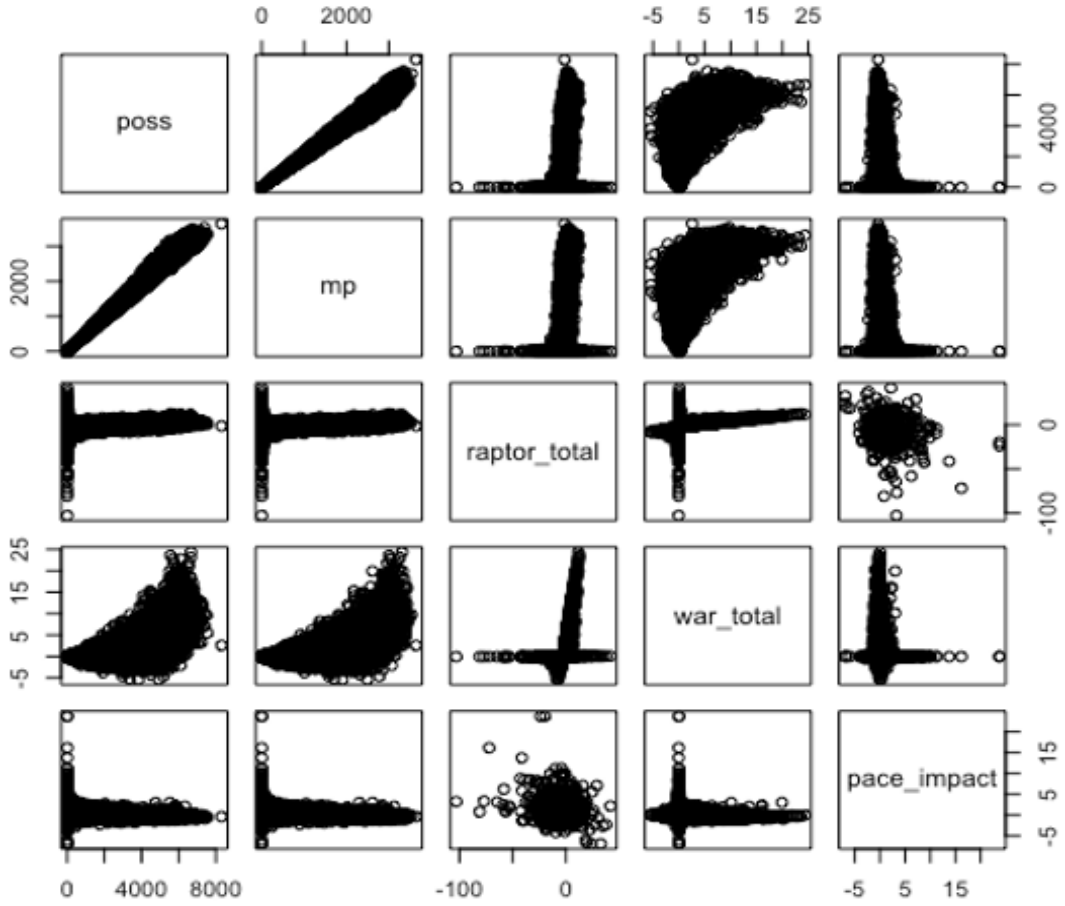
m20 <- lm(war_total ~ mp + poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_RS).

In order to further improve the model, we also check the N-folder for this first-order regression model. The correlation between the prediction and actual sets is 0.777, and it is a good

number for this model which means that the model fits well with this set of data. Also, we want to check for multicollinearity for this model.

Multicollinearity

	poss	mp	raptor_total	war_total	pace_impact
poss	1.000	0.995	0.509	0.725	-0.388
mp	0.995	1.000	0.514	0.732	-0.392
raptor_total	0.509	0.514	1.000	0.594	-0.344
war_total	0.725	0.732	0.594	1.000	-0.265
pace_impact	-0.388	-0.392	-0.344	-0.265	1.000



From the plot of these five variables, it is clear to see the model is multicollinearity. However, we are planning to do nothing because this first-order regression model has the best adjust R-square so far, and we will continue to find the best second-order. For now, the final first-order regression model for the regular season is

m20 <- lm(war_total ~ mp + poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_RS).

Play-offs

The second part is by using first-order regression models to find the relationship between the response variable, war_total, and the explanatory variables, raptor_total and pace_impact for play-offs.

m15 <- lm(war_total ~ raptor_total, data = RAPTOR_CLEAN_PO)

```
Residuals:
   Min     1Q Median     3Q    Max
-6.820 -0.346 -0.237  0.097  6.107

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45600    0.00889    51.3   <2e-16 ***
raptor_total  0.04101    0.00113    36.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.741 on 7028 degrees of freedom
Multiple R-squared:  0.157,     Adjusted R-squared:  0.157
F-statistic: 1.31e+03 on 1 and 7028 DF,  p-value: <2e-16
```

m16 <- lm(war_total ~ pace_impact, data = RAPTOR_CLEAN_PO)

```
Residuals:
   Min     1Q Median     3Q    Max
-1.743 -0.437 -0.247  0.121  6.538

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45169    0.00968    46.7   <2e-16 ***
pace_impact -0.10611    0.00717   -14.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.795 on 7028 degrees of freedom
Multiple R-squared:  0.0302,    Adjusted R-squared:  0.0301
F-statistic:  219 on 1 and 7028 DF,  p-value: <2e-16
```

m17 <- lm(war_total ~ raptor_total + pace_impact, data = RAPTOR_CLEAN_PO)

```
Residuals:
   Min     1Q Median     3Q    Max
-6.454 -0.347 -0.235  0.095  6.112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.46608    0.00901   51.71  < 2e-16 ***
raptor_total  0.03901    0.00118   33.17  < 2e-16 ***
pace_impact  -0.04311    0.00693   -6.22  5.4e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.739 on 7027 degrees of freedom
Multiple R-squared:  0.161,     Adjusted R-squared:  0.161
F-statistic:  677 on 2 and 7027 DF,  p-value: <2e-16
```

After deriving the above three results of regression models, we can find that these three F-value are pretty good. By comparing their adjusted R-squared, although the value of m17 is the largest of the three models, 16.1% is too small. At this stage, our team decided by adding two more explanatory variables, mp and poss, to find a better model.

m21 <- lm(war_total ~ mp + poss + raptor_total, data = RAPTOR_CLEAN_PO)

```
Residuals:
   Min     1Q Median     3Q    Max
-2.741 -0.185  0.022  0.150  3.850

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.161409   0.007655  -21.09   <2e-16 ***
mp            0.004450   0.000314   14.18   <2e-16 ***
poss         -0.000692   0.000162   -4.27    2e-05 ***
raptor_total  0.018982   0.000705   26.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.443 on 7026 degrees of freedom
Multiple R-squared:  0.699,     Adjusted R-squared:  0.699
F-statistic: 5.45e+03 on 3 and 7026 DF,  p-value: <2e-16
```

m22 <- lm(war_total ~ mp + poss + pace_impact, data = RAPTOR_CLEAN_PO)

```
Residuals:
   Min     1Q Median     3Q    Max
-2.873 -0.198  0.037  0.187  3.974

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.211938   0.008059  -26.30  < 2e-16 ***
mp           0.004633   0.000329   14.07  < 2e-16 ***
poss        -0.000686   0.000170   -4.03  5.6e-05 ***
pace_impact -0.009091   0.004275   -2.13    0.033 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.465 on 7026 degrees of freedom
Multiple R-squared:  0.669,     Adjusted R-squared:  0.668
F-statistic: 4.72e+03 on 3 and 7026 DF,  p-value: <2e-16
```

m23 <- lm(war_total ~ mp + poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_PO)

```
Residuals:
   Min     1Q Median     3Q    Max
-2.869 -0.181  0.024  0.149  3.838

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.168467   0.007831  -21.51  < 2e-16 ***
mp            0.004457   0.000313   14.22  < 2e-16 ***
poss         -0.000688   0.000162   -4.25  2.1e-05 ***
raptor_total  0.019685   0.000724   27.18  < 2e-16 ***
pace_impact   0.017415   0.004182    4.16  3.2e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.442 on 7025 degrees of freedom
Multiple R-squared:   0.7,      Adjusted R-squared:   0.7
F-statistic: 4.1e+03 on 4 and 7025 DF,  p-value: <2e-16
```

After adding two more explanatory variables, comparing three adjusted R-squared, m21 and m23 are very similar. Therefore, we chose the m23 regression model. The F-value is pretty good, and each terms' P-values are extremely small. For double check the last regression model for the play-offs is the best one at this stage, we ran backward and forward selection processes.

Backward

```
Start:  AIC=-11464
war_total ~ poss + mp + raptor_total + pace_impact

               Df Sum of Sq  RSS    AIC
<none>                      1374 -11464
- pace_impact   1       3.4 1378 -11449
- poss          1       3.5 1378 -11448
- mp            1      39.6 1414 -11267
- raptor_total  1     144.5 1519 -10764
> step$anova #display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
war_total ~ poss + mp + raptor_total + pace_impact

Final Model:
war_total ~ poss + mp + raptor_total + pace_impact


  Step Df Deviance Resid. Df Resid. Dev    AIC
1                      7025       1374 -11464
```

Forward

```
Start:  AIC=-3006
war_total ~ 1

               Df Sum of Sq  RSS    AIC
+ mp            1      3059 1523 -10747
+ poss          1      3020 1563 -10567
+ raptor_total  1       719 3864  -4204
+ pace_impact   1       138 4444  -3220
<none>                      4582  -3006

Step:  AIC=-10747
war_total ~ mp

               Df Sum of Sq  RSS    AIC
+ raptor_total  1     142.0 1381 -11433
+ poss          1       3.5 1520 -10761
+ pace_impact   1       1.0 1522 -10749
<none>                      1523 -10747

Step:  AIC=-11433
war_total ~ mp + raptor_total

              Df Sum of Sq  RSS    AIC
+ poss         1      3.58 1378 -11449
+ pace_impact  1      3.43 1378 -11448
<none>                     1381 -11433

Step:  AIC=-11449
war_total ~ mp + raptor_total + poss
```

```
              Df Sum of Sq  RSS    AIC
+ pace_impact  1      3.39 1374 -11464
<none>                     1378 -11449

Step:  AIC=-11464
war_total ~ mp + raptor_total + poss + pace_impact
```

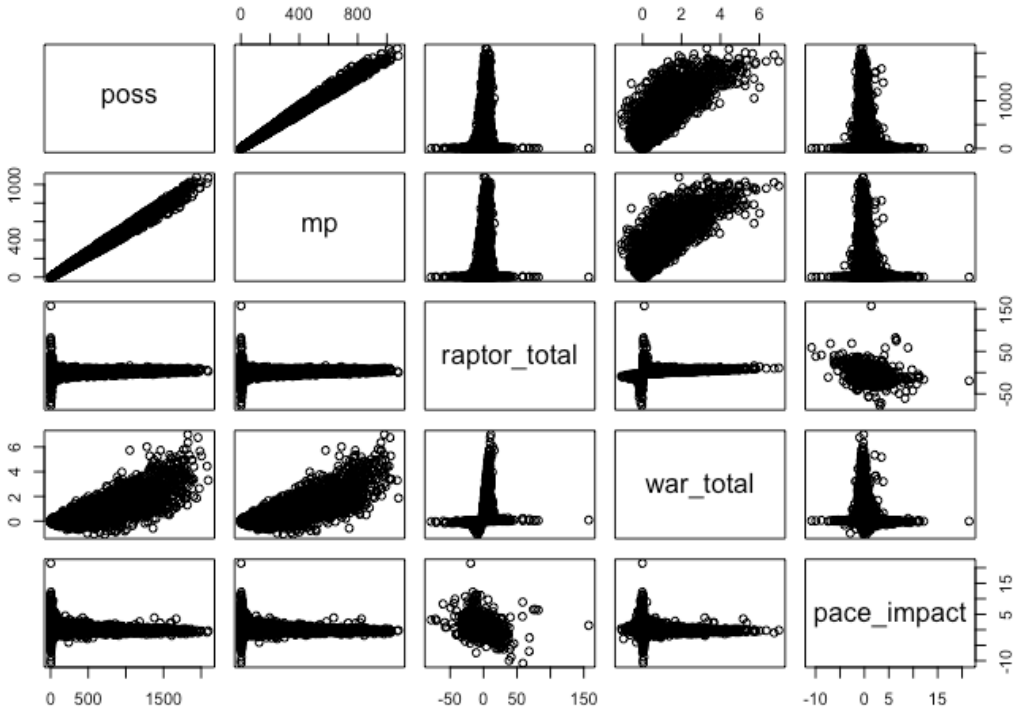
After ran backward and forward selection processes, the final results are the same, which means we are in the right direction. Therefore, for the moment, we set the final model for the play-offs as

m23 <- lm(war_total ~ mp + poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_PO).

In order to further improve the model, we also check the N-folder for this first-order regression model. The correlation between prediction set and actual set is 0.835 and it is a good number for this model which means that the model fits well with this set of data. Also, we want to check for multicollinearity for this model.

Multicollinearity

	poss	mp	raptor_total	war_total	pace_impact
poss	1.000	0.996	0.277	0.812	-0.195
mp	0.996	1.000	0.278	0.817	-0.195
raptor_total	0.277	0.278	1.000	0.396	-0.274
war_total	0.812	0.817	0.396	1.000	-0.174
pace_impact	-0.195	-0.195	-0.274	-0.174	1.000



From the plot of these five variables, it is clear to see the model is multicollinearity. However, we are planning to do nothing on it because this first-order regression model has the best adjusted R-square, and we will continue to find the best second-order model. For now, the final first-order regression model for play-offs is

m23 <- lm(war_total ~ mp + poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_PO).

In sum, it appears that exploratory variables, mp, poss, raptor_total, and pace_impact, all have a high impact on the final regression model, especially mp and poss. Before adding these two variables, mp and poss, the adjusted R-square values are quite low. Even though, after adding these two variables, the R-square value is still around 70%, it is pretty good at this stage. While we ran backward and forward selection processes, the results were the same as our decision. It might be that our dataset is too large so that our final regression models have multicollinearity problems. However, multicollinearity is a very common problem in data.

4.2.2 Second-order Regression

I will dig into the details of the raptor_total in this section. By using the second-order model, we can not only study the relationship between the variables last week but also add explanatory variables, raptor_offense, and raptor_defense. This report will also be done separately for the Regular Season and Play-offs. Further exploration is warranted to determine if we should consider using only raptor_total or both raptor_offense and raptor_defense.

m24	Regular Season	war_total ~ mp + poss + I(raptor_total^2) + pace_impact
m25	Regular Season	war_total ~ mp + poss + raptor_total + I(pace_impact^2)
m26	Regular Season	war_total ~ mp + poss + raptor_total + pace_impact + raptor_total + pace_impact + raptor_total*pace_impact + I(raptor_total^2) + I(pace_impact^2)
m27	Regular Season	war_total ~ mp + poss + I(raptor_offense^2) + I(raptor_defense^2) + pace_impact
m28	Regular Season	war_total ~ mp + poss + raptor_offense + raptor_defense + I(pace_impact^2)
m29	Regular Season	war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense + raptor_offense*pace_impact + raptor_defense*pace_impact + raptor_offense*raptor_defense*pace_impact I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2)
m30	Play-Offs	war_total ~ mp + poss + I(raptor_total^2) + pace_impact
m31	Play-Offs	war_total ~ mp + poss + raptor_total + I(pace_impact^2)
m32	Play-Offs	war_total ~ mp + poss + raptor_total + pace_impact + raptor_total + pace_impact + raptor_total*pace_impact + I(raptor_total^2) + I(pace_impact^2)
m33	Play-Offs	war_total ~ mp + poss + I(raptor_offense^2) + I(raptor_defense^2) + pace_impact
m34	Play-Offs	war_total ~ mp + poss + raptor_offense + raptor_defense + I(pace_impact^2)
m35	Play-Offs	war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense + raptor_offense*pace_impact + raptor_defense*pace_impact + raptor_offense*raptor_defense*pace_impact I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2)

Regular Season

The first step is by using second-order regression models to find the relationship between the response variable, war_total, and the explanatory variables, raptor_total and pace_impact for regular season.

m24 <- lm(war_total ~ mp + poss + I(raptor_total^2) + pace_impact, data = RAPTOR_CLEAN_RS)

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.2764  -1.0540   0.1298   0.8931  17.5098

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.232e+00  2.904e-02 -42.440  < 2e-16 ***
mp                 3.398e-03  1.748e-04  19.438  < 2e-16 ***
poss              -4.949e-04  8.597e-05  -5.756 8.76e-09 ***
I(raptor_total^2)  9.095e-04  1.061e-04   8.574  < 2e-16 ***
pace_impact        6.295e-02  1.917e-02   3.284  0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.056 on 17168 degrees of freedom
Multiple R-squared:  0.5387,    Adjusted R-squared:  0.5386
F-statistic:  5013 on 4 and 17168 DF,  p-value: < 2.2e-16
```

m25 <- lm(war_total ~ mp + poss + I(raptor_total^2) + pace_impact, data = RAPTOR_CLEAN_RS)

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.2764  -1.0540   0.1298   0.8931  17.5098

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.232e+00  2.904e-02 -42.440  < 2e-16 ***
mp                 3.398e-03  1.748e-04  19.438  < 2e-16 ***
poss              -4.949e-04  8.597e-05  -5.756 8.76e-09 ***
I(raptor_total^2)  9.095e-04  1.061e-04   8.574  < 2e-16 ***
pace_impact        6.295e-02  1.917e-02   3.284  0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.056 on 17168 degrees of freedom
Multiple R-squared:  0.5387,    Adjusted R-squared:  0.5386
F-statistic:  5013 on 4 and 17168 DF,  p-value: < 2.2e-16
```

m26 <- lm(war_total ~ mp + poss + raptor_total + pace_impact + raptor_total + pace_impact + raptor_total*pace_impact + I(raptor_total^2) + I(pace_impact^2), data = RAPTOR_CLEAN_RS)

```
Residuals:
     Min       1Q   Median       3Q      Max
-24.0978  -0.8755  -0.1954   0.6810  14.2756

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.518e-02  3.027e-02   0.501    0.616
mp                        2.333e-03  1.515e-04  15.397  < 2e-16 ***
poss                     -2.961e-04  7.423e-05  -3.989 6.67e-05 ***
raptor_total              3.141e-01  4.095e-03  76.696  < 2e-16 ***
pace_impact               1.709e-01  2.048e-02   8.343  < 2e-16 ***
I(raptor_total^2)         4.777e-03  1.203e-04  39.724  < 2e-16 ***
I(pace_impact^2)         -1.764e-02  2.681e-03  -6.581 4.80e-11 ***
raptor_total:pace_impact -1.568e-02  1.377e-03 -11.380  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.774 on 17165 degrees of freedom
Multiple R-squared:  0.6567,    Adjusted R-squared:  0.6566
F-statistic:  4692 on 7 and 17165 DF,  p-value: < 2.2e-16
```

After deriving the above three results of regression models, we can find that these three adjusted R^2 values are all not good. We decided to use a complete second-order model with two more explanatory variables, raptor_offense, and raptor_defense, instead of raptor_total.

m27 <- lm(war_total ~ mp + poss + I(raptor_offense^2) + I(raptor_defense^2) + pace_impact, data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7665 -0.1824  0.0249  0.1492  3.8370

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.1723797  0.0078332 -22.006  < 2e-16 ***
mp                0.0044538  0.0003129  14.232  < 2e-16 ***
poss             -0.0006820  0.0001615  -4.223 2.44e-05 ***
raptor_total      0.0191678  0.0007039  27.231  < 2e-16 ***
I(pace_impact^2)  0.0036680  0.0005870   6.249 4.38e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4416 on 7025 degrees of freedom
Multiple R-squared:  0.701,     Adjusted R-squared:  0.7008
F-statistic:  4118 on 4 and 7025 DF,  p-value: < 2.2e-16
```

m28 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + I(pace_impact^2), data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.9898 -0.9788 -0.1066  0.6598 19.4978

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.284e-01  2.970e-02  -7.691 1.54e-14 ***
mp                2.571e-03  1.620e-04  15.875  < 2e-16 ***
poss             -3.419e-04  7.956e-05  -4.297 1.74e-05 ***
raptor_offense    1.800e-01  4.525e-03  39.783  < 2e-16 ***
raptor_defense    2.298e-01  7.110e-03  32.316  < 2e-16 ***
I(pace_impact^2)  2.565e-02  1.951e-03  13.147  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.902 on 17167 degrees of freedom
Multiple R-squared:  0.6053,    Adjusted R-squared:  0.6052
F-statistic:  5265 on 5 and 17167 DF,  p-value: < 2.2e-16
```

m29 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense + raptor_offense*pace_impact + raptor_defense*pace_impact + raptor_offense*raptor_defense*pace_impact + I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2), data = RAPTOR_CLEAN_RS)

```
Residuals:
    Min      1Q  Median      3Q     Max
-32.453  -0.833  -0.201   0.659  13.601

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               1.326e-01  2.995e-02   4.429 9.54e-06 ***
mp                                        2.188e-03  1.476e-04  14.832  < 2e-16 ***
poss                                     -2.566e-04  7.226e-05  -3.551 0.000385 ***
raptor_offense                            3.373e-01  4.976e-03  67.781  < 2e-16 ***
raptor_defense                            3.920e-01  7.434e-03  52.737  < 2e-16 ***
pace_impact                               8.544e-02  2.068e-02   4.132 3.62e-05 ***
I(raptor_offense^2)                       4.201e-03  1.959e-04  21.440  < 2e-16 ***
I(raptor_defense^2)                       3.033e-03  3.505e-04   8.651  < 2e-16 ***
I(pace_impact^2)                         -3.342e-02  2.888e-03 -11.571  < 2e-16 ***
raptor_offense:raptor_defense             2.191e-02  6.520e-04  33.602  < 2e-16 ***
raptor_offense:pace_impact               -4.868e-02  2.057e-03 -23.670  < 2e-16 ***
raptor_defense:pace_impact               -7.451e-02  3.489e-03 -21.355  < 2e-16 ***
raptor_offense:raptor_defense:pace_impact -4.694e-03  1.656e-04 -28.349  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.726 on 17160 degrees of freedom
Multiple R-squared:  0.6751,    Adjusted R-squared:  0.6749
F-statistic:  2971 on 12 and 17160 DF,  p-value: < 2.2e-16
```

After adding two more explanatory variables and comparing three adjusted R-squared, the last regression model, m29, is the best which is 67.49%. Also, the F-value is pretty good, and each terms' P-values are extremely small. For double check the last regression model for the regular season is the best one at this stage, we ran backward and forward selection processes.

Backward

```
Start:  AIC=18762.77
war_total ~ poss + mp + raptor_offense + raptor_defense + pace_impact +
    raptor_offensedefense + raptor_offensepace + raptor_offensesqft +
    raptor_defensesqft + pace_impactsqft + opd + raptor_defensepace

                        Df Sum of Sq   RSS   AIC
<none>                                51131 18763
- poss                   1      37.6 51169 18773
- pace_impact            1      50.9 51182 18778
- raptor_defensesqft     1     223.0 51354 18836
- pace_impactsqft        1     398.9 51530 18894
- mp                     1     655.5 51787 18980
- raptor_defensepace     1    1358.9 52490 19211
- raptor_offensesqft     1    1369.7 52501 19215
- raptor_offensepace     1    1669.4 52801 19312
- opd                    1    2394.7 53526 19547
- raptor_offensedefense  1    3364.4 54496 19855
- raptor_defense         1    8286.9 59418 21340
- raptor_offense         1   13689.4 64821 22835
> step$anova #display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
war_total ~ poss + mp + raptor_offense + raptor_defense + pace_impact +
    raptor_offensedefense + raptor_offensepace + raptor_offensesqft +
    raptor_defensesqft + pace_impactsqft + opd + raptor_defensepace

Final Model:
war_total ~ poss + mp + raptor_offense + raptor_defense + pace_impact +
    raptor_offensedefense + raptor_offensepace + raptor_offensesqft +
    raptor_defensesqft + pace_impactsqft + opd + raptor_defensepace


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                     17160   51131.38 18762.77
```

Forward

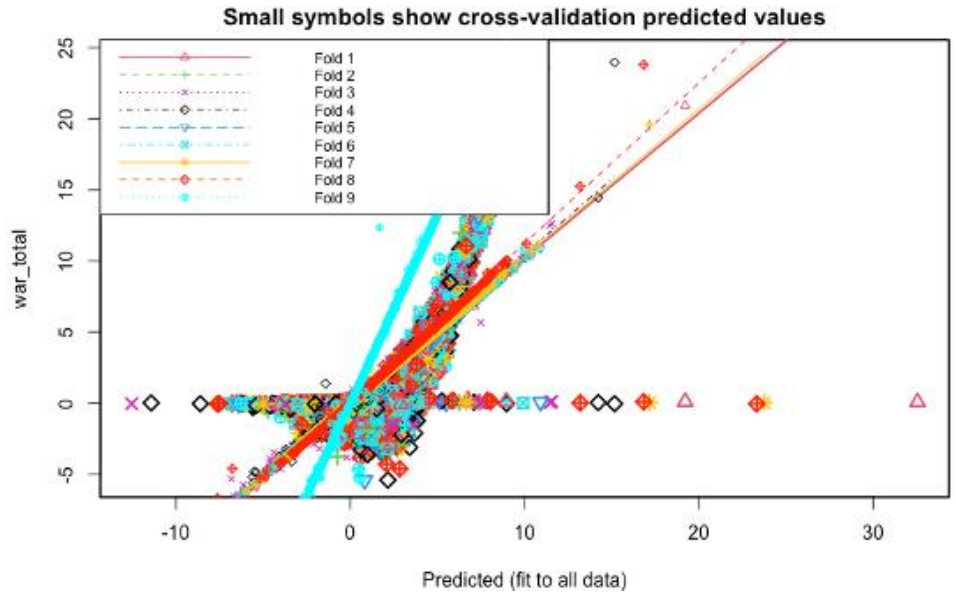
```
       Df Sum of Sq   RSS   AIC
+ poss  1    37.572 51131 18763
<none>              51169 18773

Step:  AIC=18762.77
war_total ~ mp + raptor_offense + raptor_defense + raptor_offensedefense +
    raptor_offensesqft + raptor_defensesqft + raptor_offensepace +
    opd + raptor_defensepace + pace_impactsqft + pace_impact +
    poss
```

After running backward and forward selection processes, the final results are the same, which means we are in the right direction. Therefore, at this point, we set the final model for the regular season as

m29 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense + raptor_offense*pace_impact + raptor_defense*pace_impact + raptor_offense*raptor_defense*pace_impact + I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2), data = RAPTOR_CLEAN_RS).

In order to further improve the model, we also check the K-folder for this second-order regression model. The correlation between the prediction and actual sets is 0.8376002, and it is a good number for this model which means that the model fits well with this set of data. Also, we want to check for multicollinearity for this model.

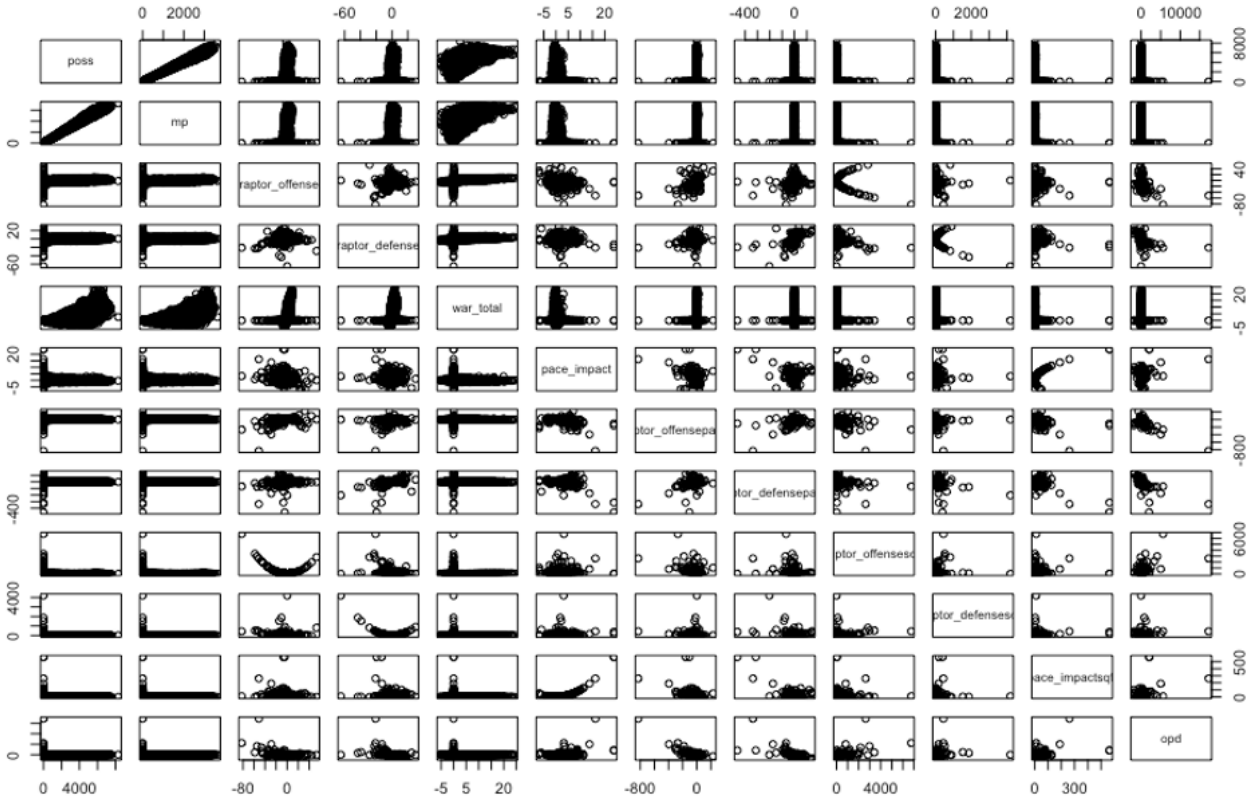
K-Folder



Multicollinearity

	poss	mp	raptor_offense	raptor_defense	war_total	pace_impact	raptor_offensedefense	raptor_offensepace
poss	1.0000	0.9954	0.495	0.259	0.7255	-0.388	-0.0920	0.1379
mp	0.9954	1.0000	0.498	0.264	0.7316	-0.392	-0.0927	0.1391
raptor_offense	0.4947	0.4984	1.000	0.237	0.5437	-0.364	-0.4448	0.4059
raptor_defense	0.2595	0.2639	0.237	1.000	0.3631	-0.124	-0.3475	0.1594
war_total	0.7255	0.7316	0.544	0.363	1.0000	-0.265	-0.0408	0.0704
pace_impact	-0.3882	-0.3923	-0.364	-0.124	-0.2651	1.000	0.1476	-0.4733
raptor_offensedefense	-0.0920	-0.0927	-0.445	-0.347	-0.0408	0.148	1.0000	-0.4249
raptor_offensepace	0.1379	0.1391	0.406	0.159	0.0704	-0.473	-0.4249	1.0000
raptor_offensesqft	-0.1422	-0.1430	-0.441	-0.237	-0.0575	0.189	0.5824	-0.5069
raptor_defensesqft	-0.0756	-0.0757	-0.107	-0.298	-0.0267	0.119	0.1615	-0.1501
pace_impactsqft	-0.1100	-0.1106	-0.147	-0.112	-0.0518	0.577	0.1503	-0.5134
opd	-0.0356	-0.0358	-0.244	-0.227	-0.0183	0.241	0.5881	-0.7265
raptor_defensepace	0.0524	0.0524	0.162	0.378	0.0253	-0.336	-0.3807	0.4408

	raptor_offensesqft	raptor_defensesqft	pace_impactsqft	opd	raptor_defensepace
poss	-0.1422	-0.0756	-0.1100	-0.0356	0.0524
mp	-0.1430	-0.0757	-0.1106	-0.0358	0.0524
raptor_offense	-0.4413	-0.1071	-0.1475	-0.2445	0.1619
raptor_defense	-0.2371	-0.2982	-0.1119	-0.2271	0.3785
war_total	-0.0575	-0.0267	-0.0518	-0.0183	0.0253
pace_impact	0.1887	0.1190	0.5774	0.2411	-0.3358
raptor_offensedefense	0.5824	0.1615	0.1503	0.5881	-0.3807
raptor_offensepace	-0.5069	-0.1501	-0.5134	-0.7265	0.4408
raptor_offensesqft	1.0000	0.2207	0.1814	0.4927	-0.2328
raptor_defensesqft	0.2207	1.0000	0.1411	0.1479	-0.3061
pace_impactsqft	0.1814	0.1411	1.0000	0.4104	-0.6488
opd	0.4927	0.1479	0.4104	1.0000	-0.6382
raptor_defensepace	-0.2328	-0.3061	-0.6488	-0.6382	1.0000



From the plot of these 12 variables, it is clear to see the model is multicollinearity. However, we are planning to do nothing because this second-order regression model has the best adjust R-square so far, and there are more than tens of thousands of data for this dataset. We will group discuss later to compare all the regression model to decide the best model for this dataset. For now, the final second-order regression model for the regular season is

m29 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense + raptor_offense*pace_impact + raptor_defense*pace_impact + raptor_offense*raptor_defense*pace_impact + I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2), data = RAPTOR_CLEAN_RS)

Play-offs

The second part is by using second-order regression models to find the relationship between the response variable, war_total, and the explanatory variables, raptor_total, pace_impact, raptor_offense, and raptor_defense for play-offs.

m30 <- lm(war_total ~ mp + poss + I(raptor_total^2) + pace_impact, data = RAPTOR_CLEAN_PO)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.8779 -0.1972  0.0369  0.1872  3.9526

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.194e-01  8.151e-03 -26.920  < 2e-16 ***
mp                 4.634e-03  3.287e-04  14.098  < 2e-16 ***
poss              -6.781e-04  1.697e-04  -3.997 6.49e-05 ***
I(raptor_total^2)  7.998e-05  1.425e-05   5.612 2.07e-08 ***
pace_impact       -1.059e-02  4.274e-03  -2.477   0.0133 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4639 on 7025 degrees of freedom
Multiple R-squared:   0.67,     Adjusted R-squared:  0.6698
F-statistic:  3566 on 4 and 7025 DF,  p-value: < 2.2e-16
```

m31 <- lm(war_total ~ mp + poss + raptor_total + I(pace_impact^2), data = RAPTOR_CLEAN_PO)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7665 -0.1824  0.0249  0.1492  3.8370

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.1723797  0.0078332 -22.006  < 2e-16 ***
mp                 0.0044538  0.0003129  14.232  < 2e-16 ***
poss              -0.0006820  0.0001615  -4.223 2.44e-05 ***
raptor_total       0.0191678  0.0007039  27.231  < 2e-16 ***
I(pace_impact^2)  0.0036680  0.0005870   6.249 4.38e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4416 on 7025 degrees of freedom
Multiple R-squared:  0.701,     Adjusted R-squared:  0.7008
F-statistic:  4118 on 4 and 7025 DF,  p-value: < 2.2e-16
```

m32 <- lm(war_total ~ mp + poss + raptor_total + pace_impact + raptor_total + pace_impact + raptor_total*pace_impact + I(raptor_total^2) + I(pace_impact^2), data = RAPTOR_CLEAN_PO)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.6893 -0.1792  0.0255  0.1451  3.8244

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.716e-01  8.005e-03 -21.440  < 2e-16 ***
mp                        4.440e-03  3.120e-04  14.232  < 2e-16 ***
poss                     -6.790e-04  1.610e-04  -4.218 2.50e-05 ***
raptor_total              2.095e-02  7.715e-04  27.161  < 2e-16 ***
pace_impact               1.144e-02  4.882e-03   2.343   0.0191 *
I(raptor_total^2)        -3.319e-06  1.394e-05  -0.238   0.8118
I(pace_impact^2)          6.785e-04  7.832e-04   0.866   0.3863
raptor_total:pace_impact -1.563e-03  2.309e-04  -6.772 1.37e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4402 on 7022 degrees of freedom
Multiple R-squared:  0.703,     Adjusted R-squared:  0.7027
F-statistic:  2375 on 7 and 7022 DF,  p-value: < 2.2e-16
```

After deriving the above three results of regression models, we can find that these three adjusted R^2 values are all acceptable. However, we decided to use a complete second-order model with two more explanatory variables, raptor_offense, and raptor_defense, instead of raptor_total.

m33 <- lm(war_total ~ mp + poss + I(raptor_offense^2) + I(raptor_defense^2) + pace_impact, data = RAPTOR_CLEAN_PO)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.8822 -0.1959  0.0361  0.1898  3.9370

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.267e-01  8.314e-03 -27.270  < 2e-16 ***
mp                   4.637e-03  3.283e-04  14.126  < 2e-16 ***
poss                -6.717e-04  1.694e-04  -3.964 7.44e-05 ***
I(raptor_offense^2)  2.088e-04  3.510e-05   5.948 2.84e-09 ***
I(raptor_defense^2)  7.962e-05  3.786e-05   2.103   0.0355 *
pace_impact         -9.044e-03  4.308e-03  -2.099   0.0358 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4633 on 7024 degrees of freedom
Multiple R-squared:  0.6709,    Adjusted R-squared:  0.6707
F-statistic:  2864 on 5 and 7024 DF,  p-value: < 2.2e-16
```

m34 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + I(pace_impact^2), data = RAPTOR_CLEAN_PO)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7355 -0.1830  0.0251  0.1490  3.8351

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.1719674  0.0078572 -21.887  < 2e-16 ***
mp                0.0044612  0.0003131  14.246  < 2e-16 ***
poss             -0.0006862  0.0001616  -4.245 2.21e-05 ***
raptor_offense    0.0195794  0.0009304  21.045  < 2e-16 ***
raptor_defense    0.0183685  0.0013752  13.357  < 2e-16 ***
I(pace_impact^2)  0.0036851  0.0005876   6.272 3.78e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4416 on 7024 degrees of freedom
Multiple R-squared:  0.701,     Adjusted R-squared:  0.7008
F-statistic:  3294 on 5 and 7024 DF,  p-value: < 2.2e-16
```

m35 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense + raptor_offense*pace_impact + raptor_defense*pace_impact + raptor_offense*raptor_defense*pace_impact + I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2), data = RAPTOR_CLEAN_PO)

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.2378 -0.1778  0.0242  0.1533  3.7895

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.793e-01  8.189e-03 -21.890  < 2e-16 ***
mp                                       4.394e-03  3.094e-04  14.204  < 2e-16 ***
poss                                    -6.546e-04  1.596e-04  -4.102 4.15e-05 ***
raptor_offense                           2.139e-02  9.987e-04  21.417  < 2e-16 ***
raptor_defense                           2.777e-02  1.719e-03  16.158  < 2e-16 ***
pace_impact                              1.954e-02  5.215e-03   3.747  0.00018 ***
I(raptor_offense^2)                      4.752e-04  5.515e-05   8.616  < 2e-16 ***
I(raptor_defense^2)                     -2.071e-04  5.241e-05  -3.952 7.83e-05 ***
I(pace_impact^2)                         3.713e-03  8.304e-04   4.471 7.90e-06 ***
raptor_offense:raptor_defense           -2.290e-04  1.027e-04  -2.229  0.02585 *
raptor_offense:pace_impact               2.357e-03  4.418e-04   5.335 9.87e-08 ***
raptor_defense:pace_impact              -3.542e-03  4.518e-04  -7.839 5.20e-15 ***
raptor_offense:raptor_defense:pace_impact -4.904e-05  3.704e-05  -1.324  0.18550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.436 on 7017 degrees of freedom
Multiple R-squared:  0.7089,    Adjusted R-squared:  0.7084
F-statistic:  1424 on 12 and 7017 DF,  p-value: < 2.2e-16
```

Since the P-value for term raptor_offense * raptor_defense * pace_impact is not significant, we decided to delete the term, therefore, the new model for m35 will be:

m36 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense + raptor_offense*pace_impact + raptor_defense*pace_impact + I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2), data = RAPTOR_CLEAN_PO)

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1907 -0.1784  0.0236  0.1529  3.7921

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.784e-01  8.161e-03 -21.854  < 2e-16 ***
mp                            4.392e-03  3.094e-04  14.195  < 2e-16 ***
poss                         -6.542e-04  1.596e-04  -4.099 4.19e-05 ***
raptor_offense                2.132e-02  9.973e-04  21.376  < 2e-16 ***
raptor_defense                2.808e-02  1.703e-03  16.495  < 2e-16 ***
pace_impact                   1.947e-02  5.215e-03   3.733 0.000191 ***
I(raptor_offense^2)           4.732e-04  5.513e-05   8.583  < 2e-16 ***
I(raptor_defense^2)          -2.224e-04  5.111e-05  -4.352 1.37e-05 ***
I(pace_impact^2)              3.577e-03  8.241e-04   4.340 1.44e-05 ***
raptor_offense:raptor_defense -2.917e-04  9.117e-05  -3.200 0.001383 **
raptor_offense:pace_impact    2.423e-03  4.390e-04   5.519 3.52e-08 ***
raptor_defense:pace_impact   -3.573e-03  4.512e-04  -7.919 2.76e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.436 on 7018 degrees of freedom
Multiple R-squared:  0.7088,    Adjusted R-squared:  0.7084
F-statistic:  1553 on 11 and 7018 DF,  p-value: < 2.2e-16
```

After adding two more explanatory variables and comparing three adjusted R-squared, the last regression model, m36, is the best which is 70.84%. Also, the F-value is pretty good, and each terms' P-values are extremely small. For double check the last regression model for the regular season is the best one at this stage, we ran backward and forward selection processes.

Backward

```
Start:  AIC=-11655.35
war_total ~ poss + mp + raptor_offense + raptor_defense + pace_impact +
    raptor_offensepace + raptor_defensepace + raptor_offensesqft +
    raptor_defensesqft + pace_impactsqft + opd

                    Df Sum of Sq    RSS    AIC
<none>                           1334.9 -11655
- opd                1     1.335 1336.2 -11650
- pace_impact        1     2.243 1337.1 -11646
- poss               1     3.203 1338.1 -11640
- pace_impactsqft    1     4.035 1338.9 -11636
- raptor_defensesqft 1     4.352 1339.2 -11634
- raptor_offensepace 1     4.465 1339.3 -11634
- raptor_defensepace 1    10.766 1345.6 -11601
- raptor_offensesqft 1    14.338 1349.2 -11582
- mp                 1    38.315 1373.2 -11458
- raptor_defense     1    49.831 1384.7 -11400
- raptor_offense     1    86.449 1421.3 -11216
```

Forward

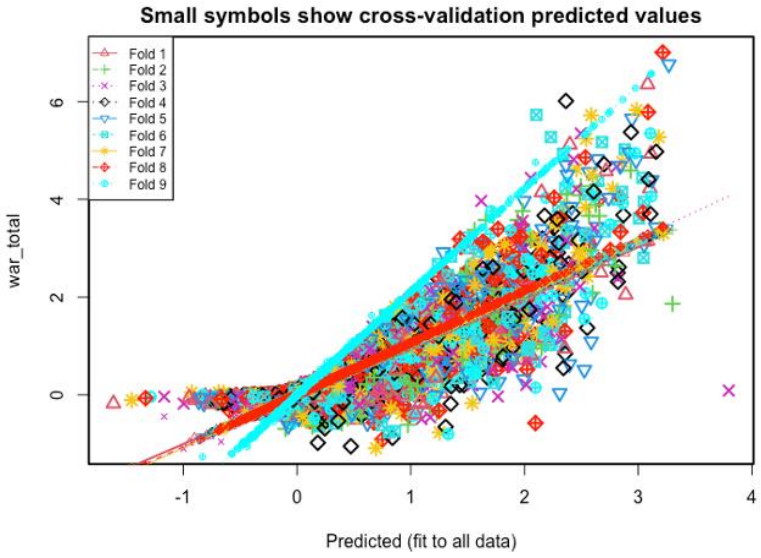
```
       Df Sum of Sq    RSS    AIC
+ opd   1    1.3351 1334.9 -11655
<none>              1336.2 -11650

Step:  AIC=-11655.35
war_total ~ mp + raptor_offense + raptor_defense + raptor_defensepace +
    raptor_offensesqft + raptor_defensesqft + pace_impact + poss +
    raptor_offensepace + pace_impactsqft + opd
```

After running backward and forward selection processes, the final results for forward and backward are not the same as our result. However, after compared to the adjust R-square of forward/backward result and our result, at this point, we decided to set the final model for the play-offs as

m36 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_offense*raptor_defense*pace_impact + raptor_offense*pace_impact + raptor_defense*pace_impact + I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2), data = RAPTOR_CLEAN_PO)

In order to further improve the model, we also check the K-folder for this second-order regression model. The correlation between prediction set and the actual set is 0.8419767 and it is a good number for this model which means that the model fits well with this set of data. Also, we want to check for multicollinearity for this model.

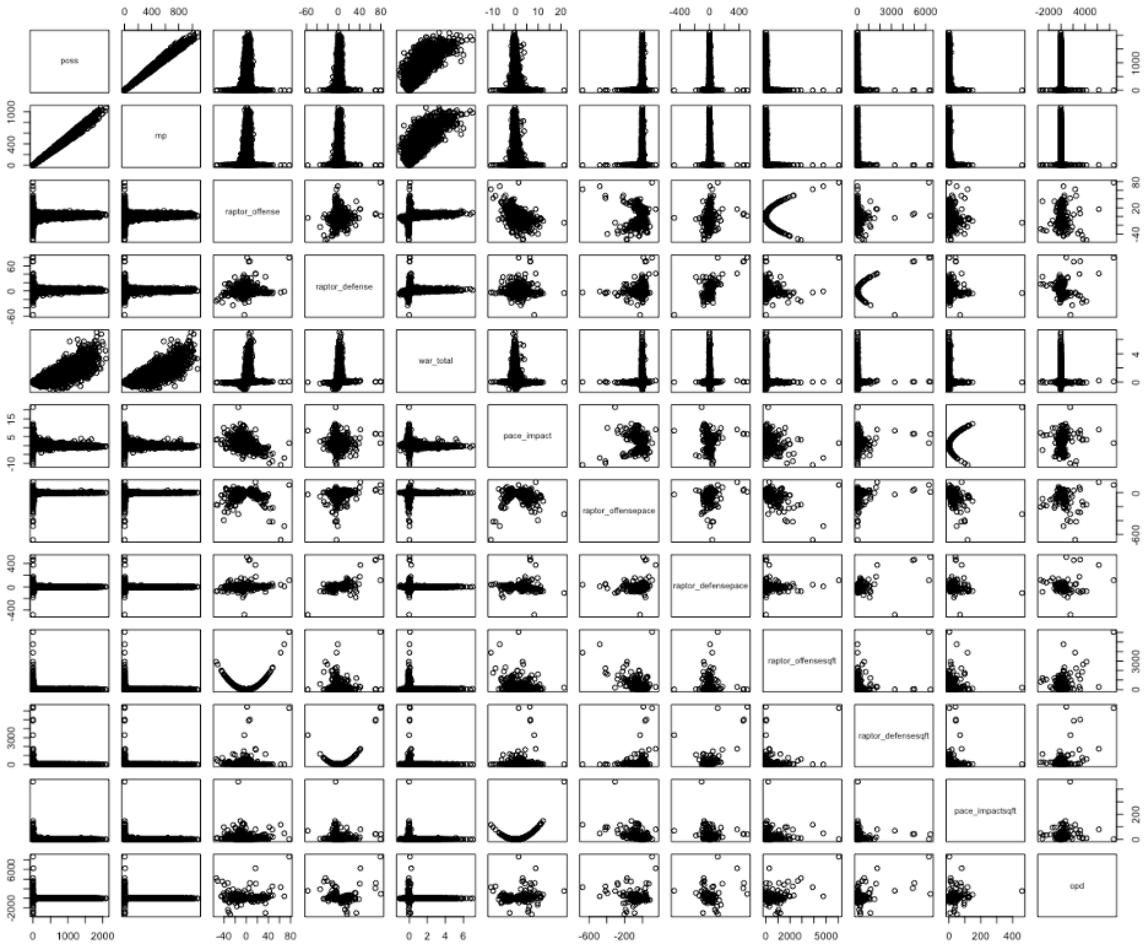
K-Folder



Multicollinearity

	poss	mp	raptor_offense	raptor_defense	war_total	pace_impact	raptor_offensepace
poss	1.00000	0.99640	0.2618	0.15147	0.81177	-0.1952	0.1157
mp	0.99640	1.00000	0.2610	0.15438	0.81705	-0.1954	0.1153
raptor_offense	0.26180	0.26098	1.0000	0.20540	0.36001	-0.3703	0.0333
raptor_defense	0.15147	0.15438	0.2054	1.00000	0.23848	0.0187	0.1278
war_total	0.81177	0.81705	0.3600	0.23848	1.00000	-0.1738	0.0745
pace_impact	-0.19520	-0.19542	-0.3703	0.01866	-0.17383	1.0000	-0.2071
raptor_offensepace	0.11569	0.11529	0.0333	0.12782	0.07447	-0.2071	1.0000
raptor_defensepace	-0.00984	-0.00984	0.1184	0.53205	0.00112	-0.0353	0.1387
raptor_offensesqft	-0.15172	-0.15135	-0.0180	0.00689	-0.07670	-0.0147	-0.5977
raptor_defensesqft	-0.06996	-0.06967	0.0568	0.47027	-0.03197	0.1344	0.0331
pace_impactsqft	-0.14866	-0.14838	-0.0946	-0.01785	-0.08805	0.5107	-0.5721
opd	-0.04840	-0.04834	0.0329	0.07027	-0.02911	0.1464	-0.2069

	raptor_defensepace	raptor_offensesqft	raptor_defensesqft	pace_impactsqft	opd
poss	-0.00984	-0.15172	-0.0700	-0.1487	-0.0484
mp	-0.00984	-0.15135	-0.0697	-0.1484	-0.0483
raptor_offense	0.11840	-0.01804	0.0568	-0.0946	0.0329
raptor_defense	0.53205	0.00689	0.4703	-0.0178	0.0703
war_total	0.00112	-0.07670	-0.0320	-0.0881	-0.0291
pace_impact	-0.03534	-0.01468	0.1344	0.5107	0.1464
raptor_offensepace	0.13866	-0.59769	0.0331	-0.5721	-0.2069
raptor_defensepace	1.00000	0.01171	0.4801	-0.0823	0.0920
raptor_offensesqft	0.01171	1.00000	0.2782	0.2503	0.4663
raptor_defensesqft	0.48015	0.27822	1.0000	0.1418	0.5045
pace_impactsqft	-0.08227	0.25025	0.1418	1.0000	0.2562
opd	0.09199	0.46632	0.5045	0.2562	1.0000



From the plot of these 12 variables, it is clear to see the model is multicollinearity. However, we are planning to do nothing because this second-order regression model has the best adjust R-square so far, and there are more than tens of thousands of data for this dataset. We will group discuss later to compare all the regression models to decide the best model for this dataset. For now, the final second-order regression model for the play-offs season is

```
m35 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact +
raptor_offense*raptor_defense*pace_impact + raptor_offense*pace_impact +
raptor_defense*pace_impact  + I(raptor_offense^2) + I(raptor_defense^2) + I(pace_impact^2),
data = RAPTOR_CLEAN_PO).
```

In sum, it appears that exploratory variables, mp, poss, raptor_total, and pace_impact, all have a high impact on the final regression model, especially mp and poss. Before adding these two variables, mp and poss, the adjusted R-square values are quite low. We discovered that even though the effect of raptor_offense and raptor_defense are bigger than raptor_total, the effect is not too big. After adding these two variables, the R-square value is still around 70%; it is pretty good. While we ran backward and forward selection processes, the results were the same as our decision for Regular Season. Although the result of backward and forward selection for Play-Offs is not the same as ours, the difference between these two is not distinct. It might be that our dataset is too large so that our final regression models have multicollinearity problems. However, multicollinearity is a very common problem in data.

Subsection 4.3: Agraj Allola

4.3.1 First-order Regression

NBA is a huge league with multiple players joining every year and each player has their own skillset, this makes it difficult to have a fair evaluation for the players.

In this report we are trying to evaluate a variable to determine the evaluation of the player using two variables called war_total and composite_elo. These are dependent variables which dependent on various independent variables to find a fair evaluation for the players.

We have started with a dataset called historical_raptor_byteam.This dataset consists of various independent variables like raptor_offense,raptor_defense, etc. to determine the war_total of the player.war_total is an independent variable used as an evaluation. To add another dimension to themodel we have added another dataset called composite_elo. This contains the elo rating of all players present in the league. We have merged it into a single dataset.

We have removed duplicate variables which were present in both datasets like player_name,player_id etc.We have removed elo data from 2015 to 2019 as they were inconsistent.We have used all the variables in our model up to this point as they play a role in the analysis. We used correlation matrix to check multi-collinearity.

```
M28
Call:
lm(formula = war_total ~ mp, data = RAPTOR_CLEAN_RS_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-10.1427  -1.0715   0.1206   0.9221  17.6975

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.134e+00  2.540e-02  -44.63   <2e-16 ***
mp           2.353e-03  1.673e-05  140.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.064 on 17171 degrees of freedom
Multiple R-squared:  0.5353,      Adjusted R-squared:  0.5353
F-statistic: 1.978e+04 on 1 and 17171 DF,  p-value: < 2.2e-16
```

From above model we can say that mp is dependent on war_total. Its variability is 5%. F value is 1.978e+04. Adjusted R squared of this model is 0.5353. t value of team_rank is 140.64. P-value is less than 0.05 so this model is fit.

M29
Call:
lm(formula = war_total ~ poss, data = RAPTOR_CLEAN_RS_1_)

Residuals:
Min 1Q Median 3Q Max
-9.9939 -1.0899 0.1166 0.8961 18.2732

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.088e+00 2.552e-02 -42.64 <2e-16 ***
poss 1.150e-03 8.325e-06 138.13 <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.084 on 17171 degrees of freedom
Multiple R-squared: 0.5263, Adjusted R-squared: 0.5263
F-statistic: 1.908e+04 on 1 and 17171 DF, p-value: < 2.2e-16

From above model we can say that poss is dependent on war_total. Its variability is 5%. F value is 1.908e+04. Adjusted R squared of this model is 0.5263. t value of team_rank is -42.64. P-value is less than 0.05 so this model is fit.

M30
Call:
lm(formula = war_total ~ poss + mp, data = RAPTOR_CLEAN_RS_1_)

Residuals:
Min 1Q Median 3Q Max
-10.2288 -1.0624 0.1205 0.9250 17.6891

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.137e+00 2.538e-02 -44.808 < 2e-16 ***
poss -4.863e-04 8.618e-05 -5.643 1.69e-08 ***
mp 3.336e-03 1.749e-04 19.074 < 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.062 on 17170 degrees of freedom
Multiple R-squared: 0.5362, Adjusted R-squared: 0.5361
F-statistic: 9924 on 2 and 17170 DF, p-value: < 2.2e-16

From above model we can say that poss and mp are dependent on war_total. Its variability is 5%. F value is 9924. Adjusted R squared of this model is 0.52. t value is -5.643 and 19.074. P-value is less than 0.05 so this model is fit.

M34
Call:
lm(formula = war_total ~ mp + raptor_offense + raptor_defense +
pace_impact, data = RAPTOR_CLEAN_RS_1_)

Residuals:
Min 1Q Median 3Q Max
-8.5265 -0.9842 -0.1138 0.6678 19.5924

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.375e-01 3.108e-02 -10.86 <2e-16 ***
mp 1.942e-03 1.869e-05 103.88 <2e-16 ***
raptor_offense 1.882e-01 4.604e-03 40.89 <2e-16 ***
raptor_defense 2.235e-01 7.084e-03 31.55 <2e-16 ***
pace_impact 2.551e-01 1.788e-02 14.27 <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.902 on 17168 degrees of freedom
Multiple R-squared: 0.6056, Adjusted R-squared: 0.6055
F-statistic: 6589 on 4 and 17168 DF, p-value: < 2.2e-16

From above model we can say that raptor_defense,raptor_offense and mp are dependent on war_total. Its variability is 6%. F value is 6589. Adjusted R squared of this model is 0.605. t value 103.88,40.89,31.55,14.27. P-value is less than 0.05 so this model is fit.

M35
Call:
lm(formula = war_total ~ poss + raptor_offense + raptor_defense +
pace_impact, data = RAPTOR_CLEAN_RS_1_)

Residuals:
Min 1Q Median 3Q Max
-8.3689 -0.9910 -0.1316 0.6629 19.9893

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.703e-01 3.102e-02 -8.715 <2e-16 ***
poss 9.413e-04 9.238e-06 101.895 <2e-16 ***
raptor_offense 1.920e-01 4.632e-03 41.463 <2e-16 ***

```
raptor_defense  2.286e-01  7.131e-03  32.055   <2e-16 ***
pace_impact    2.412e-01  1.799e-02  13.406   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.916 on 17168 degrees of freedom
Multiple R-squared:  0.5997,            Adjusted R-squared:  0.5996
F-statistic:  6430 on 4 and 17168 DF,  p-value: < 2.2e-16
```

From above model we can say that raptor_defense,raptor_offense and poss are dependent on war_total. Its variability is 5.9%. F value is 6430. Adjusted R squared of this model is 0.5996. t value are 101.895,41.463,32.055. P-value is less than 0.05 so this model is fit.

M38

```
Call:
lm(formula = war_total ~ mp + raptor_total + pace_impact, data = RAPTOR_CLEAN_RS_1_)

Residuals:
      Min      1Q  Median      3Q     Max
-8.5001 -0.9832 -0.1167  0.6634 19.9752

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.282e-01  3.100e-02  -10.59   <2e-16 ***
mp                1.937e-03  1.866e-05  103.80   <2e-16 ***
raptor_total  1.994e-01  3.632e-03   54.91   <2e-16 ***
pace_impact   2.630e-01  1.778e-02   14.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.902 on 17169 degrees of freedom
Multiple R-squared:  0.6052,            Adjusted R-squared:  0.6051
F-statistic:  8773 on 3 and 17169 DF,  p-value: < 2.2e-16
```

From above model we can say that raptor_total,pace_impact and mp are dependent on war_total. Its variability is 6%. F value is 8773. Adjusted R squared of this model is 0.6051 . t value are 103.80,54.91,14.79. P-value is less than 0.05 so this model is fit.

M39
Call:
lm(formula = war_total ~ poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_RS_1_)

Residuals:
Min 1Q Median 3Q Max
-8.7141 -0.9926 -0.1303 0.6543 20.3866

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.605e-01 3.094e-02 -8.421 <2e-16 ***
poss 9.387e-04 9.221e-06 101.806 <2e-16 ***
raptor_total 2.037e-01 3.649e-03 55.807 <2e-16 ***
pace_impact 2.493e-01 1.789e-02 13.938 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.916 on 17169 degrees of freedom
Multiple R-squared: 0.5993, Adjusted R-squared: 0.5993
F-statistic: 8561 on 3 and 17169 DF, p-value: < 2.2e-16

From above model we can say that raptor_total,pace_impact and poss are dependent on war_total. Its variability is 5.9%. F value is 8561. Adjusted R squared of this model is 0.5993 . t value are 103.80,54.91,14.79. P-value is less than 0.05 so this model is fit.

M31
Call:
lm(formula = war_total ~ mp, data = RAPTOR_CLEAN_PO)

Residuals:
Min 1Q Median 3Q Max
-2.8256 -0.2014 0.0369 0.1904 4.0220

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.182e-01 7.743e-03 -28.18 <2e-16 ***
mp 3.321e-03 2.796e-05 118.80 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4656 on 7028 degrees of freedom
Multiple R-squared: 0.6676, Adjusted R-squared: 0.6675
F-statistic: 1.411e+04 on 1 and 7028 DF, p-value: < 2.2e-16

```
M32
Call:
lm(formula = war_total ~ poss, data = RAPTOR_CLEAN_PO)

Residuals:
      Min      1Q  Median      3Q     Max
-2.6803 -0.2018  0.0358  0.1897  4.1538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.156e-01  7.852e-03  -27.46   <2e-16 ***
poss         1.703e-03  1.462e-05  116.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4715 on 7028 degrees of freedom
Multiple R-squared:  0.659, Adjusted R-squared:  0.6589
F-statistic: 1.358e+04 on 1 and 7028 DF,  p-value: < 2.2e-16

M33
Call:
lm(formula = war_total ~ poss + mp, data = RAPTOR_CLEAN_PO)

Residuals:
      Min      1Q  Median      3Q     Max
-2.8765 -0.1982  0.0370  0.1901  3.9743

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2166921  0.0077442 -27.981  < 2e-16 ***
poss        -0.0006835  0.0001701  -4.019  5.9e-05 ***
mp           0.0046408  0.0003295  14.086  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4651 on 7027 degrees of freedom
Multiple R-squared:  0.6683,    Adjusted R-squared:  0.6682
F-statistic:  7080 on 2 and 7027 DF,  p-value: < 2.2e-16

M36
Call:
lm(formula = war_total ~ mp + raptor_offense + raptor_defense +
    pace_impact, data = RAPTOR_CLEAN_PO)

Residuals:
      Min      1Q  Median      3Q     Max
```

```
-2.8107 -0.1817  0.0225  0.1495  3.8859

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.698e-01  7.834e-03  -21.67   <2e-16 ***
mp              3.129e-03  2.792e-05  112.10   <2e-16 ***
raptor_offense  2.067e-02  9.943e-04   20.79   <2e-16 ***
raptor_defense  1.797e-02  1.388e-03   12.95   <2e-16 ***
pace_impact     1.925e-02  4.355e-03  4.42     1e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4428 on 7025 degrees of freedom
Multiple R-squared:  0.6994,            Adjusted R-squared:  0.6992
F-statistic:  4086 on 4 and 7025 DF,  p-value: < 2.2e-16
```

M37

```
Call:
lm(formula = war_total ~ poss + raptor_offense + raptor_defense +
    pace_impact, data = RAPTOR_CLEAN_PO)

Residuals:
      Min      1Q  Median      3Q     Max
-2.8620 -0.1838  0.0203  0.1502  4.0351

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.664e-01  7.945e-03 -20.941  < 2e-16 ***
poss            1.603e-03  1.459e-05 109.843  < 2e-16 ***
raptor_offense  2.063e-02  1.008e-03  20.472  < 2e-16 ***
raptor_defense  1.863e-02  1.405e-03  13.259  < 2e-16 ***
pace_impact     1.834e-02  4.411e-03   4.159 3.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4486 on 7025 degrees of freedom
Multiple R-squared:  0.6915,            Adjusted R-squared:  0.6913
F-statistic:  3937 on 4 and 7025 DF,  p-value: < 2.2e-16
```

M40

```
Call:
lm(formula = war_total ~ mp + raptor_total + pace_impact, data = RAPTOR_CLEAN_PO)

Residuals:
      Min      1Q  Median      3Q     Max
```

```
-2.8673 -0.1825  0.0234  0.1500  3.8905

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.701e-01  7.832e-03 -21.715  < 2e-16 ***
mp            3.130e-03  2.792e-05 112.105  < 2e-16 ***
raptor_total  1.968e-02  7.252e-04  27.142  < 2e-16 ***
pace_impact   1.751e-02  4.187e-03   4.183 2.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4428 on 7026 degrees of freedom
Multiple R-squared:  0.6993,    Adjusted R-squared:  0.6992
F-statistic:  5447 on 3 and 7026 DF,  p-value: < 2.2e-16

M41
Call:
lm(formula = war_total ~ poss + raptor_total + pace_impact, data = RAPTOR_CLEAN_PO)

Residuals:
     Min      1Q  Median      3Q     Max
-2.9038 -0.1837  0.0208  0.1511  4.0385

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.666e-01  7.942e-03 -20.981  < 2e-16 ***
poss          1.603e-03  1.459e-05 109.856  < 2e-16 ***
raptor_total  1.990e-02  7.345e-04  27.091  < 2e-16 ***
pace_impact   1.706e-02  4.242e-03   4.023 5.82e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4486 on 7026 degrees of freedom
Multiple R-squared:  0.6915,    Adjusted R-squared:  0.6913
F-statistic:  5248 on 3 and 7026 DF,  p-value: < 2.2e-16
```

Since all the values of p are less than 0.05 so there is strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

Backward Selection

```
Start:  AIC=22057.5
war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact

               Df Sum of Sq   RSS   AIC
<none>                       61996 22058
- poss          1      79.4 62075 22078
- pace_impact   1     747.3 62743 22261
- mp            1     999.7 62996 22330
- raptor_defense  1    3559.8 65556 23014
- raptor_offense  1    6034.4 68030 23651

Call:
lm(formula = war_total ~ mp + poss + raptor_offense + raptor_defense +
    pace_impact, data = RAPTOR_CLEAN_RS_1_)

Coefficients:
   (Intercept)            mp          poss  raptor_offense
    -0.3438820      0.0026962     -0.0003727      0.1880773
raptor_defense    pace_impact
     0.2224083      0.2571162
```

Forward Selection

```
Start:  AIC=22057.5
war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact

Call:
lm(formula = war_total ~ mp + poss + raptor_offense + raptor_defense +
    pace_impact, data = RAPTOR_CLEAN_RS_1_)

Coefficients:
   (Intercept)            mp          poss  raptor_offense
    -0.3438820      0.0026962     -0.0003727      0.1880773
raptor_defense    pace_impact
     0.2224083      0.2571162
```

We have completed mulitcolinearity test in module 7 and we have created a correlation matrix to check the multicolinearity if the variables.The figure is given below.

Residual Analysis for m34 and m36

M34

Using the resid() function we have viewed the residuals present in the model m34.The figure shows the residuals of the model.

- Using the qqnorm() function we have plotted the residuals in a graph.Using this we can observe that the residuals follow a linear path.So it follows normal distribution.
- We have also used qqline() to add a structure to the plot.
- We observe that the plot is linear and it does not have any clustered resdiuals.We also can observe that it follows normal disribution.There are no prominent outliers.

M36

Using the resid() function we have viewed the residuals present in the model m36.The figure shows the residuals of the model.

- Similarily,using the qqnorm() function we have plotted the residuals in a graph.Using this we can observe that the residuals follow a linear path.So it follows normal distribution.
- We have also used qqline() to add a structure to the plot.
- We observe that the plot is linear but not as linear as M34 and it does not have any clustered resdiuals.We also can observe that it follows normal disribution.There are no prominent outliers but we can see some outliers.

The plots show that 4 assumptions of residuals are positive in both the cases. This shows that this model has a high predictability in the equation.

4.3.2 Second-order Regression

Regular season

```
m1 <- lm(war_total~ poss + (raptor_total^2), data = RAPTOR_CLEAN_RS_1_)
Call:
lm(formula = war_total ~ poss + (raptor_total^2), data = RAPTOR_CLEAN_RS_1_)

Residuals:
   Min     1Q  Median     3Q    Max
-8.3847 -1.0034 -0.1172  0.6473 20.0991

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.271e-01  2.959e-02  -4.295 1.75e-05 ***
```

```
poss          9.048e-04  8.944e-06 101.165  < 2e-16 ***
raptor_total  1.943e-01  3.607e-03  53.863  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.927 on 17170 degrees of freedom
Multiple R-squared:  0.5948,  Adjusted R-squared:  0.5948
F-statistic: 1.26e+04 on 2 and 17170 DF,  p-value: < 2.2e-16


m2 <- lm(war_total~ mp + poss + (raptor_total^2), data = RAPTOR_CLEAN_RS_1_)

Call:
lm(formula = war_total ~ mp + poss + (raptor_total^2), data = RAPTOR_CLEAN_RS_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6203 -0.9881 -0.1068  0.6578 19.6447

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.901e-01  2.964e-02  -6.414 1.46e-10 ***
mp            2.579e-03  1.629e-04  15.832  < 2e-16 ***
poss         -3.541e-04  8.001e-05  -4.425 9.69e-06 ***
raptor_total  1.893e-01  3.595e-03  52.645  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.913 on 17169 degrees of freedom
Multiple R-squared:  0.6006,  Adjusted R-squared:  0.6006
F-statistic:  8607 on 3 and 17169 DF,  p-value: < 2.2e-16


m3 <- lm(war_total~ mp + poss + (raptor_total^2)+ (pace_impact^2), data =
RAPTOR_CLEAN_RS_1_)

Call:
lm(formula = war_total ~ mp + poss + (raptor_total^2) + (pace_impact^2),
    data = RAPTOR_CLEAN_RS_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5561 -0.9838 -0.1187  0.6625 19.9277
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.350e-01  3.101e-02 -10.800  < 2e-16 ***
mp            2.706e-03  1.621e-04  16.690  < 2e-16 ***
poss         -3.796e-04  7.952e-05  -4.774 1.82e-06 ***
raptor_total  1.990e-01  3.631e-03  54.798  < 2e-16 ***
pace_impact   2.648e-01  1.777e-02  14.903  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.901 on 17168 degrees of freedom
Multiple R-squared:  0.6057,  Adjusted R-squared:  0.6056
F-statistic:  6594 on 4 and 17168 DF,  p-value: < 2.2e-16


m4 <- lm(war_total~ mp + poss + (raptor_total^2)+ (pace_impact^2) + raptor_offense +
raptor_defense, data = RAPTOR_CLEAN_RS_1_)
Call:
lm(formula = war_total ~ mp + poss + (raptor_total^2) + (pace_impact^2) +
    raptor_offense + raptor_defense, data = RAPTOR_CLEAN_RS_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5653 -0.9833 -0.1158  0.6672 19.5560

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.3438820  0.0310885 -11.061  < 2e-16 ***
mp              0.0026962  0.0001621  16.638  < 2e-16 ***
poss           -0.0003727  0.0000795  -4.688 2.78e-06 ***
raptor_total    0.2224083  0.0070839  31.396  < 2e-16 ***
pace_impact     0.2571162  0.0178735  14.385  < 2e-16 ***
raptor_offense -0.0343310  0.0089112  -3.853 0.000117 ***
raptor_defense         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.9 on 17167 degrees of freedom
Multiple R-squared:  0.6061,  Adjusted R-squared:  0.606
F-statistic:  5282 on 5 and 17167 DF,  p-value: < 2.2e-16
```

Playoff Season

```
M5<- lm(war_total~ poss + (raptor_total^2), data = RAPTOR_CLEAN_PO)
Call:
lm(formula = war_total ~ poss + (raptor_total^2), data = RAPTOR_CLEAN_PO)

Residuals:
   Min      1Q  Median      3Q     Max
-2.7783 -0.1878  0.0202  0.1517  4.0500

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.597e-01  7.762e-03  -20.58   <2e-16 ***
poss          1.596e-03  1.449e-05  110.14   <2e-16 ***
raptor_total  1.921e-02  7.150e-04   26.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4491 on 7027 degrees of freedom
Multiple R-squared:  0.6907,  Adjusted R-squared:  0.6907
F-statistic:  7848 on 2 and 7027 DF,  p-value: < 2.2e-16

m6 <- lm(war_total~mp +  poss + (raptor_total^2), data = RAPTOR_CLEAN_PO)
Call:
lm(formula = war_total ~ mp + poss + (raptor_total^2), data = RAPTOR_CLEAN_PO)

Residuals:
   Min      1Q  Median      3Q     Max
-2.7409 -0.1846  0.0223  0.1496  3.8499

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1614086  0.0076546 -21.087  < 2e-16 ***
mp            0.0044497  0.0003138  14.180  < 2e-16 ***
poss         -0.0006917  0.0001619  -4.272 1.97e-05 ***
raptor_total  0.0189819  0.0007052  26.918  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4428 on 7026 degrees of freedom
Multiple R-squared:  0.6993,  Adjusted R-squared:  0.6992
F-statistic:  5448 on 3 and 7026 DF,  p-value: < 2.2e-16
```

Backward selection

```
Start:  AIC=22057.5
war_total ~ mp + poss + (raptor_total^2) + (pace_impact^2) +
    raptor_offense + raptor_defense


Step:  AIC=22057.5
war_total ~ mp + poss + raptor_total + pace_impact + raptor_offense

                Df Sum of Sq   RSS   AIC
<none>                       61996 22058
- raptor_offense  1      53.6 62050 22070
- poss            1      79.4 62075 22078
- pace_impact     1     747.3 62743 22261
- mp              1     999.7 62996 22330
- raptor_total    1    3559.8 65556 23014

Call:
lm(formula = war_total ~ mp + poss + raptor_total + pace_impact +
    raptor_offense, data = RAPTOR_CLEAN_RS_1_)

Coefficients:
   (Intercept)             mp           poss    raptor_total
    -0.3438820       0.0026962      -0.0003727       0.2224083
   pace_impact  raptor_offense
     0.2571162      -0.0343310
```

Forward selection

```
Start:  AIC=22057.5
war_total ~ mp + poss + (raptor_total^2) + (pace_impact^2) +
    raptor_offense + raptor_defense


Call:
lm(formula = war_total ~ mp + poss + (raptor_total^2) + (pace_impact^2) +
    raptor_offense + raptor_defense, data = RAPTOR_CLEAN_RS_1_)

Coefficients:
   (Intercept)             mp           poss    raptor_total
```

```
 -0.3438820       0.0026962      -0.0003727       0.2224083
pace_impact  raptor_offense  raptor_defense
  0.2571162      -0.0343310             NA
```

From the above model we can conclude that composite_elo is not dependent on all variables. P value of war_total is more than 0.05. its variability is 17%.

The variables in the dataset had a different impact on the evaluation of the player. Before starting our calculations, we have decided to segregate the data into two parts (Playoffs, Regular season). This was done for examining the difference in values in each type of season. To obtain which variable has the most created plots using ggplot in R. To find the multicollinearity we used a function called cor (name of dataset) to find the collinearity between all sets of the variables present in the dataset. Using these plots and by reviewing them in comparison against each other we have selected variables that had more relationship (collinearity, variance) between them.

Secondly, we started the model building process. We created various permutations of sets of variables among both the response variables and the explanatory variables. I build 14 models using war_total as the response variable and a combination of the variables (mp, poss, raptor_offense, raptor_defense, raptor_total, pace_impact) as the explanatory variables)

I built the models using the linear model function in R (lm) and saved them to perform further analysis on them. I have performed a summary calculation using the summary() function in r, to view the coefficients and the intercept of the model created. Using this we can gain information on how much a variable in the model affects the 'value' of the player. We have also viewed other factors in the function like Adjusted R-squared, Multiple R squared, F-statistic, and p-value to view the impact of the model. Using these I chose the best ', first model'.

To further refine our analysis, we used second-order regression. We started by the dataset into Regular season and Playoff Season and further dividing our dataset into two sets of training dataset and testing dataset of ratio 80:20. We have used second-order terms and interaction terms on our models to add more insight. After performing a summary function on these models and factoring all the results of summary function and cor() and visual plots we decided our second order best model. To verify and re-check our second model best model we performed stepAIC (forward and backward selection) and we also performed N cross-validation to have more evidence. To test the assumptions of a linear model we have performed residual analysis, to check the autocorrelation we performed the Durbin Watson test, to check normality we plotted histograms, to check residuals we used the sum(model$residuals). We also performed n cross-validation and added more insight into our best model. Using this model, we can find the most accurate evaluation of the player.

Subsection 4.4: Pramatesh Shukla

4.4.1 First-order Regression

m42	Regular Season	team_rank ~ war_total
m43	Regular Season	composite_elo ~ war_total
m44	Play-Offs	team_rank ~ war_total
m45	Play-Offs	composite_elo ~ war_total
m46	Regular Season	team_rank ~ mp + poss + raptor_total + pace_impact + war_total
m47	Regular Season	compsite_elo ~ mp + poss + raptor_total + pace_impact + war_total
m48	Play-Offs	team_rank ~ mp + poss + raptor_total + pace_impact + war_total
m49	Play-Offs	composite_elo ~ mp + poss + raptor_total + pace_impact + war_total
m50	Regular Season	team_rank ~ mp + poss + raptor_offense + raptor_defense + pace_impact + war_total
m51	Regular Season	compsite_elo ~ mp + poss + raptor_offense + raptor_defense + pace_impact + war_total
m52	Play-Offs	team_rank ~ mp + poss + raptor_offense + raptor_defense + pace_impact + war_total
m53	Play-Offs	compsite_elo ~ mp + poss + raptor_offense + raptor_defense + pace_impact + war_total

Relationship between response variables, composite_elo and team_rank, and exploratory variables: mp, poss, raptor_total, raptor_offense, raptor_deffense, pace_impact and War_total:

*Best model for **phase 1** - First-order regression:*
Regular Season: composite_elo ~ war_total, data = RAPTOR_CLEAN_RS (m43)
Play-Offs: composite_elo ~ war_total, data = RAPTOR_CLEAN_PO (m45)

*Best model for **phase 2** - Second-order regression:*
Regular season: team_rank ~ mp + poss + raptor_total + war_total + poly(pace_impact,2), data = RAPTOR_CLEAN_RS (m46)
Play offs: composite_elo ~ mp + poss + raptor_offense + pace_impact + war_total + poly(raptor_defense, 2), data = RAPTOR_CLEAN_PO (m53)

m43<-lm(composite_elo~war_total,data= RAPTOR_CLEAN_RS)
Residuals:

Min	1Q	Median	3Q	Max
-332.04	-78.96	6.05	79.41	327.19

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1497.7461	0.9636	1554.40	<2e-16	***
war_total	9.2598	0.2787	33.22	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 110.6 on 17171 degrees of freedom
Multiple R-squared: 0.06039, Adjusted R-squared: 0.06033
F-statistic: 1104 on 1 and 17171 DF, p-value: < 2.2e-16

From this model we can say that composite_elo is dependent on war_total. Its variability is 6%. F value is 1104. Adjusted R squared of this model is 0.06033. The t-value of war_total is 33.22. P- value is less than 0.05, so this model is fit.

m45<-lm(composite_elo~war_total,data= `RAPTOR_CLEAN_PO.(1)`)
Residuals:

Min	1Q	Median	3Q	Max
-198.46	-50.87	-5.34	44.84	235.38

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1582.8911	0.9052	1748.57	<2e-16	***
war_total	29.0622	0.9932	29.26	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.23 on 7028 degrees of freedom
Multiple R-squared: 0.1086, Adjusted R-squared: 0.1085
F-statistic: 856.2 on 1 and 7028 DF, p-value: < 2.2e-16

From the above model we can say that composite_elo is dependent on war_total. Its variability is 10%. F value is 856.2. Adjusted R squared of this model is 0.1085.t the value of war_total is 29.26. P value is less than 0.05 so this model is fit.

m46<-
lm(team_rank~mp+poss+raptor_total+pace_impact+war_total,data=RAPTOR_CLEAN_RS)
Residuals:

Min	1Q	Median	3Q	Max
-958.82	-342.27	-3.45	346.65	1025.20

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)

```
(Intercept)  682.970506.74293 101.287  < 2e-16 ***
mp                -0.02857         0.03541  -0.807   0.4198
poss              0.07306          0.01724   4.238 2.27e-05 ***
raptor_total  -1.49298 0.85279  -1.751   0.0800 .
pace_impact   -8.15951         3.87492  -2.106   0.0352 *
war_total         -61.01617        1.65369 -36.897  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411.9 on 17167 degrees of freedom
Multiple R-squared:  0.09326, Adjusted R-squared:  0.09299
F-statistic: 353.1 on 5 and 17167 DF,  p-value: < 2.2e-16
```

From this model, we can say that team_rank is not dependent on all variables. F value is 353.1. Adjusted R squared of this model is 0.09299. P value of mp and raptor_total is more than 0.05. its variability is 9%.

```
vif(m46)
          mp        poss raptor_total  pace_impact    war_total
 112.383054   109.733615            1.650873           1.237995           2.536344

m46<-lm(team_rank~+poss+raptor_total+pace_impact+war_total, data=RAPTOR_CLEAN_RS)
vif(m46)
          poss raptor_total  pace_impact    war_total
          2.325612     1.650077        1.232642           2.495846
```

Interaction model for m46

```
m46<-lm(team_rank~mp+poss+raptor_total+pace_impact+war_total+poss*pace_impact,
data=RAPTOR_CLEAN_RS)
Residuals:
   Min     1Q Median         3Q        Max
-960.9 -341.9   -3.8  346.7 1029.4

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          681.467899   6.776011 100.571  < 2e-16 ***
mp                   -0.025758   0.035425  -0.727  0.46718
poss                 0.071867   0.017247   4.167  3.1e-05 ***
raptor_total         -1.839025   0.866805  -2.122  0.03388 *
pace_impact          -12.132548   4.267449  -2.843  0.00447 **
war_total            -60.153302   1.698519 -35.415  < 2e-16 ***
poss:pace_impact   0.006274   0.002825   2.221  0.02635 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411.9 on 17166 degrees of freedom
Multiple R-squared:  0.09352, Adjusted R-squared:  0.0932
```

F-statistic: 295.2 on 6 and 17166 DF, p-value: < 2.2e-16

Second order model for m46

m46<-lm(team_rank~mp+poss+raptor_total+war_total+poly(pace_impact,2), data=RAPTOR_CLEAN_RS)

Residuals:

Min	1Q	Median	3Q	Max
-958.70	-342.40	-3.59	346.70	1026.07

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	681.23155	6.52348	104.428	< 2e-16	***
mp	-0.02936	0.03542	-0.829	0.4073	
poss	0.07328	0.01724	4.250	2.15e-05	***
raptor_total	-1.46533	0.85366	-1.717	0.0861	.
war_total	-61.06664	1.65519	-36.894	< 2e-16	***
poly(pace_impact, 2)1	-981.70919	458.91666	-2.139	0.0324	*
poly(pace_impact, 2)2	301.19542	417.57749	0.721	0.4707	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411.9 on 17166 degrees of freedom
Multiple R-squared: 0.09328, Adjusted R-squared: 0.09297
F-statistic: 294.3 on 6 and 17166 DF, p-value: < 2.2e-16

m53<-lm(composite_elo~mp+poss+raptor_offense+raptor_defense+pace_impact+war_total,data = `RAPTOR_CLEAN_PO.(1)`)

Residuals:

Min	1Q	Median	3Q	Max
-173.226	-48.142	-4.381	42.357	232.118

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1565.32185	1.18494	1321.014	< 2e-16	***
mp	0.39675	0.04662	8.511	< 2e-16	***
poss	-0.12908	0.02375	-5.436	5.64e-08	***
raptor_offense	-0.55795	0.15004	-3.719	0.000202	***
raptor_defense	0.43997	0.20553	2.141	0.032337	*
pace_impact	2.90328	0.63826	4.549	5.49e-06	***
war_total	1.05583	1.74842	0.604	0.545944	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.81 on 7023 degrees of freedom
Multiple R-squared: 0.1724, Adjusted R-squared: 0.1717
F-statistic: 243.9 on 6 and 7023 DF, p-value: < 2.2e-16

From the above model we can conclude that composite_elo is not dependent on all variables. F value is 243.9. Adjusted R squared of this model is 0.1717. P value of war_total is more than 0.05. its variability is 17%.

```
vif(m53)
        mp      poss raptor_offense raptor_defense  pace_impact
  143.491695 139.751232    1.342013     1.097024     1.192865
  war_total
    3.335538
```

```
m53<-lm(composite_elo~+poss+raptor_offense+raptor_defense+pace_impact+war_total,data =
`RAPTOR_CLEAN_PO.(1)`)
```

```
vif(m53)
        poss raptor_offense raptor_defense  pace_impact       war_total
    2.993346     1.338504     1.096721     1.192133     3.241510
```

Interaction model for m53

```
m53<-
lm(composite_elo~mp+poss+raptor_offense+raptor_defense+pace_impact+war_total+poss*rapt
or_defense,data = `RAPTOR_CLEAN_PO.(1)`)

Residuals:
      Min       1Q   Median       3Q      Max
-171.999  -48.050   -4.399   42.296  231.948

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)          1.565e+03  1.185e+00 1320.841  < 2e-16 ***
mp                   3.926e-01  4.662e-02    8.420  < 2e-16 ***
poss                -1.251e-01  2.379e-02   -5.258 1.50e-07 ***
raptor_offense      -4.448e-01  1.560e-01   -2.852  0.00436 **
raptor_defense       2.301e-01  2.203e-01    1.045  0.29619
pace_impact          2.998e+00  6.390e-01    4.692 2.76e-06 ***
war_total           -2.180e+00  2.134e+00   -1.021  0.30712
poss:raptor_defense  2.811e-03  1.064e-03    2.641  0.00829 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.78 on 7022 degrees of freedom
Multiple R-squared:  0.1733,  Adjusted R-squared:  0.1724
F-statistic: 210.2 on 7 and 7022 DF,  p-value: < 2.2e-16
```

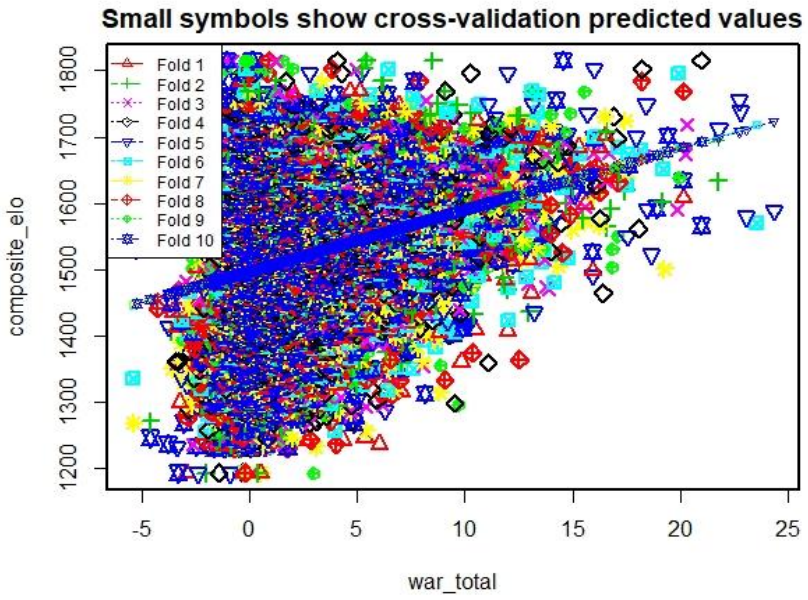
Second order model for m53

```
m53<-
lm(composite_elo~mp+poss+raptor_offense+pace_impact+war_total+poly(raptor_defense,2),dat
a = `RAPTOR_CLEAN_PO.(1)`)

Residuals:
```

```
           Min       1Q   Median       3Q      Max
-171.547  -48.137   -4.166   42.403  231.706

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   1565.43788    1.18710 1318.706  < 2e-16 ***
mp                               0.39607    0.04661    8.497  < 2e-16 ***
poss                            -0.12891    0.02374   -5.429 5.86e-08 ***
raptor_offense                  -0.54566    0.15021   -3.633 0.000283 ***
pace_impact                      3.03169    0.64297    4.715 2.46e-06 ***
war_total                        0.85999    1.75229    0.491 0.623595
poly(raptor_defense, 2)1  147.85699   67.88664    2.178 0.029439 *
poly(raptor_defense, 2)2 -108.62809   66.30093   -1.638 0.101381
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.8 on 7022 degrees of freedom
Multiple R-squared:  0.1728,  Adjusted R-squared:  0.1719
F-statistic: 209.5 on 7 and 7022 DF,  p-value: < 2.2e-16
```

Regular Season N-fold cross validation

```
m43<-lm(composite_elo~war_total,data= RAPTOR_CLEAN_RS)
summary(m43)
out <- cv.lm(data=RAPTOR_CLEAN_RS,form.lm=formula(m43),m=10)

Sum of squares = 21700894            Mean square = 12639        n = 1717

Overall (Sum over all 1717 folds)
   ms
12228
```

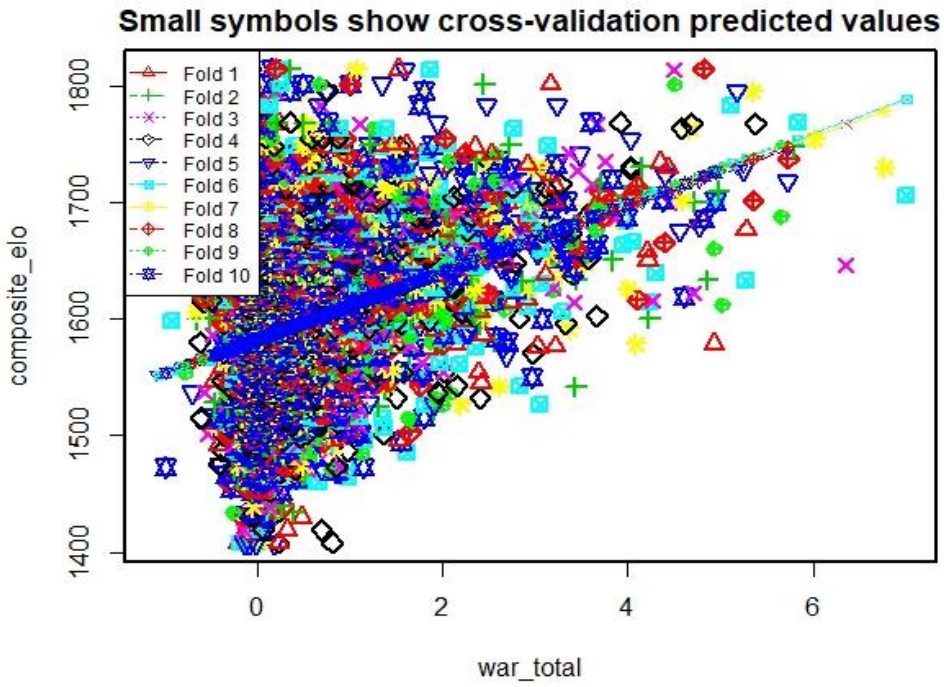


Play offs N-fold cross validation.

```
m45<-lm(composite_elo~war_total,data= `RAPTOR_CLEAN_PO.(1)`)
summary(m45)
out <- cv.lm(data=`RAPTOR_CLEAN_PO.(1)`,form.lm=formula(m45),m=10)
```

Sum of squares = 3308593Mean square = 4706 n = 703

Overall (Sum over all 703 folds)
ms
4522



In this report, we mainly focus on exploring goal two by using two new response variables, composite_elo and team_rank, and the explanatory variable, war_total. Model 46 and model53 are the best second-order models. Adjusted r-squared of m46 is 0.09297 and p-value is 2.2e-16, which is less than 0.05. The adjusted R-squared of m53 is 0.1719 and the p-value is 2.2e-16. From model 43 we can say that composite_elo is dependent on war_total. Its variability is 6%. The F-value is 1104. The adjusted R squared of this model is 0.06033. The t-value of war_total is 33.22. P-value is less than 0.05, so this model is fit. The Sum of squares of m43 is 21700894 and the mean square is 12639. From model 45 we can say that composite_elo is dependent on war_total. Its variability is 10%. The F value is 856.2. The adjusted R squared of this model is 0.1085. The t-value of war_total is 29.26. P-value is less than 0.05 so this model is fit. Sum of Squares of m45 is 3308593, Mean square is 4706. Hence, we can say that from both the model regular season and playoffs the fittest model is m45 as its adjusted R squared value is higher than m43.

Discussion

In analyzing the data, we have found that war_total or the dependent variable in the model can be used as an effective metric for determining player ‘value’. It is also continuous and quantitative. We have calculated many models with many combinations of variables, including team_rank, composite_elo, etc. The war_total variable is a combination of raptor_total, mp (minutes played), league pace, and place_impact. Although the composite Elo rating system was good, it had only three variables (peak, avg, and end) making up the calculation, which were significant. Still, they didn't consider other important variables like mp and raptor_total. Factors such as these helped give importance to our dependent variable; war_total. Each team member analyzed several different combinations of explanatory variables, after finding the best model and bringing the results together, we found that the best models contain similar explanatory variables: mp, poss, raptol_total, and pace_impact. Therefore, we were able to achieve the project's goal of NBA player ‘value’ by looking deep into the dataset for important variables like mp, poss, raptol_total, and pace_impact. Most Importantly, these variables play a critical role in determining each teams’ success.

Conclusions

The purpose of our analysis was to find the evaluation of the player using data curated in the past games. By using raw data and performing various analysis we are trying to get a well-defined model to find the evaluation. We have used datasets called "Historical_Raptor_byteam" and the Elo rating dataset. After combining these datasets, we created a new dataset consisting of the unique variables combined. Using the functions in R, we performed various analyses to find the impact of variables on each other. We have used datasets and performed analysis using to create a response variable to determine the value of the player. We had two response variables war_total and composite_elo, using our analyses, we have decided to use war_total as our response variable and the explanatory variables are mp, poss, raptol_total, and pace_impact.we first found that all of the exploratory variables significantly impact the final regression model, especially mp and poss. The Regular season had more predictability in the model than Play-offs, but the difference between those two has some minute similarities which were observed in forward and backward selections. From model 43 we can say that composite_elo is dependent on war_total. Its variability is 6% so we have chosen war_total as our response variable.
The ‘best’ model while for the play-offs in phase 2 a second-order term was chosen for the ‘best’ model which includes both raptor_offense and raptor/-defense. Finally, we have defined what objectives define the measure of the value of the player and the measure of the values which predict the team's success in our study in the report.

Appendix 4.1: Kyle Kassen

Phase 1…

Regular Season [RS] Models:

```
> RAPTOR_CLEAN_RS <- read.csv("C:/Users/IQ1006/Desktop/DSC 423/DSC 423 Group Project/RAPTOR_CL
EAN_RS.csv")
>   View(RAPTOR_CLEAN_RS)
> m1 <- lm(war_total ~ raptor_offense,data=RAPTOR_CLEAN_RS)
> summary(m1)
> m2 <- lm(war_total ~ raptor_defense,data=RAPTOR_CLEAN_RS)
> summary(m2)
> m3 <- lm(war_total ~ raptor_offense + raptor_defense,data=RAPTOR_CLEAN_RS)
> summary(m3)
> m7 <- lm(war_total ~ mp+poss+raptor_offense+pace_impact,data=RAPTOR_CLEAN_RS)
> summary (m7)
> m8 <- lm(war_total ~ mp + poss + raptor_defense + pace_impact,data=RAPTOR_CLEAN_RS)
> summary(m8)
> m9 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact,data=RAPTOR_C
LEAN_RS)
> summary(m9)
```

Model: m1

```
Call:
lm(formula = war_total ~ raptor_offense, data = RAPTOR_CLEAN_RS)

Residuals:
    Min      1Q  Median      3Q     Max
-25.676  -1.489  -0.747   0.788  33.722

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.316513   0.020837  111.17   <2e-16 ***
raptor_offense 0.439115   0.005173   84.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.541 on 17171 degrees of freedom
Multiple R-squared:  0.2956,   Adjusted R-squared:  0.2955
F-statistic:  7205 on 1 and 17171 DF,  p-value: < 2.2e-16
```

Model: m2

```
Call:
lm(formula = war_total ~ raptor_defense, data = RAPTOR_CLEAN_RS)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1282  -1.6696  -0.8676   0.8312  31.3311

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.89816    0.02199   86.33   <2e-16 ***
raptor_defense 0.51356    0.01006   51.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.821 on 17171 degrees of freedom
Multiple R-squared:  0.1319,   Adjusted R-squared:  0.1318
F-statistic:  2608 on 1 and 17171 DF,  p-value: < 2.2e-1
```

Model: m3

```
Call:
lm(formula = war_total ~ raptor_offense + raptor_defense, data = RAPTOR_CLEAN_RS)

Residuals:
    Min      1Q  Median      3Q     Max
-18.785  -1.355  -0.820   0.687  37.068

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.40327    0.02008  119.68   <2e-16 ***
raptor_offense 0.39161    0.00510   76.78   <2e-16 ***
raptor_defense 0.35115    0.00893   39.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.434 on 17170 degrees of freedom
Multiple R-squared:  0.3538,   Adjusted R-squared:  0.3537
F-statistic:  4700 on 2 and 17170 DF,  p-value: < 2.2e-16
```

Model: m7

```
Call:
lm(formula = war_total ~ mp + poss + raptor_offense + pace_impact,
    data = RAPTOR_CLEAN_RS)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4366  -1.0260  -0.0473   0.7434  16.5608

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.362e-01  3.134e-02 -17.107  < 2e-16 ***
mp              2.963e-03  1.664e-04  17.804  < 2e-16 ***
poss           -4.556e-04  8.171e-05  -5.576  2.5e-08 ***
raptor_offense  2.059e-01  4.695e-03  43.864  < 2e-16 ***
pace_impact     2.597e-01  1.838e-02  14.128  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.954 on 17168 degrees of freedom
Multiple R-squared:  0.5835,   Adjusted R-squared:  0.5834
F-statistic:  6012 on 4 and 17168 DF,  p-value: < 2.2e-16
```

Model: m8

```
Call:
lm(formula = war_total ~ mp + poss + raptor_defense + pace_impact,
    data = RAPTOR_CLEAN_RS)

Residuals:
   Min     1Q Median     3Q    Max
-9.732 -1.050  0.018  0.821 17.274

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -9.128e-01  2.912e-02 -31.346  < 2e-16 ***
mp              3.039e-03  1.695e-04  17.929  < 2e-16 ***
poss           -3.969e-04  8.328e-05  -4.766 1.89e-06 ***
raptor_defense  2.582e-01  7.364e-03  35.065  < 2e-16 ***
pace_impact     1.041e-01  1.831e-02   5.688 1.31e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.991 on 17168 degrees of freedom
Multiple R-squared:  0.5677,   Adjusted R-squared:  0.5676
F-statistic:  5637 on 4 and 17168 DF,  p-value: < 2.2e-16
```

Model: m9

```
Call:
lm(formula = war_total ~ mp + poss + raptor_offense + raptor_defense +
    pace_impact, data = RAPTOR_CLEAN_RS)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5653 -0.9833 -0.1158  0.6672 19.5560

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.3438820  0.0310885 -11.061  < 2e-16 ***
mp              0.0026962  0.0001621  16.638  < 2e-16 ***
poss           -0.0003727  0.0000795  -4.688 2.78e-06 ***
raptor_offense  0.1880773  0.0046010  40.877  < 2e-16 ***
raptor_defense  0.2224083  0.0070839  31.396  < 2e-16 ***
pace_impact     0.2571162  0.0178735  14.385  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.9 on 17167 degrees of freedom
Multiple R-squared:  0.6061,   Adjusted R-squared:  0.606
F-statistic:  5282 on 5 and 17167 DF,  p-value: < 2.2e-16
```

Regular Season stepAIC

```
> install.packages("MASS")
> library(MASS)
> m9 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact,data=RAPTOR_C
LEAN_RS)
> step <- stepAIC(m9,direction="backward")
> model_empty <- lm(war_total ~ 1, data=RAPTOR_CLEAN_RS)
> step <- stepAIC(model_empty,direction="forward",scope=list(upper=m9,lower=model_empty))
```

"backward" stepAIC

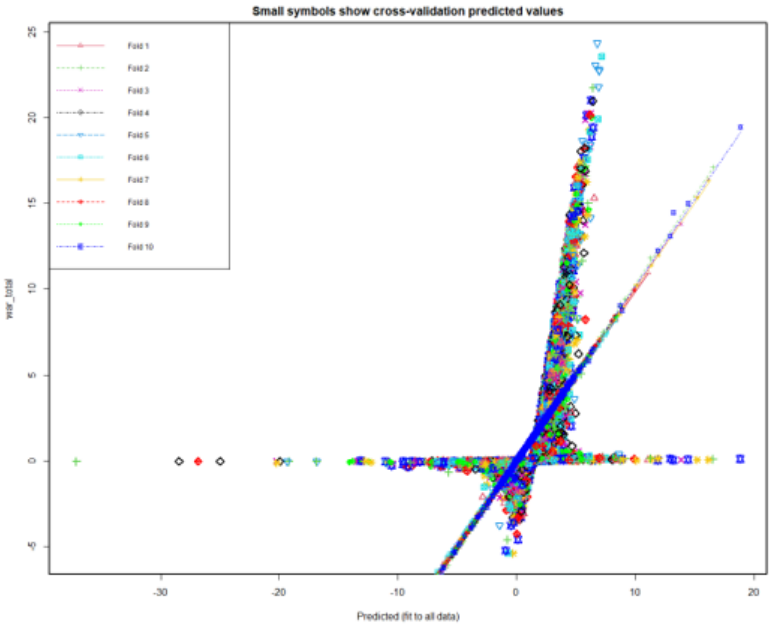
```
Start:  AIC=22057.5
war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact

                 Df Sum of Sq   RSS   AIC
<none>                       61996 22058
- poss            1      79.4 62075 22078
- pace_impact     1     747.3 62743 22261
- mp              1     999.7 62996 22330
- raptor_defense  1    3559.8 65556 23014
- raptor_offense  1    6034.4 68030 23651
```

"forward" stepAIC

```
Start:  AIC=38045.66
war_total ~ 1

                 Df Sum of Sq    RSS   AIC
+ mp              1     84245  73134 24887
+ poss            1     82834  74545 25215
+ raptor_offense  1     46516 110863 32031
+ raptor_defense  1     20753 136626 35619
+ pace_impact     1     11060 146319 36796
<none>                       157379 38046

Step:  AIC=24886.83
war_total ~ mp

                 Df Sum of Sq   RSS   AIC
+ raptor_offense  1    6711.0 66423 23236
+ raptor_defense  1    4890.7 68243 23700
+ poss            1     135.4 72998 24857
+ pace_impact     1      89.7 73044 24868
<none>                       73134 24887

Step:  AIC=23235.93
war_total ~ mp + raptor_offense

                 Df Sum of Sq   RSS   AIC
+ raptor_defense  1    3611.3 62811 22278
+ pace_impact     1     748.3 65674 23043
+ poss            1     104.9 66318 23211
<none>                       66423 23236

Step:  AIC=22277.92
war_total ~ mp + raptor_offense + raptor_defense

              Df Sum of Sq   RSS   AIC
+ pace_impact  1    736.12 62075 22078
+ poss         1     68.17 62743 22261
<none>                     62811 22278

Step:  AIC=22077.47
war_total ~ mp + raptor_offense + raptor_defense + pace_impact

       Df Sum of Sq   RSS   AIC
+ poss  1    79.368 61996 22058
<none>              62075 22078

Step:  AIC=22057.5
war_total ~ mp + raptor_offense + raptor_defense + pace_impact +
    poss
```

Regular Season N-Fold Cross-Validation

```
> install.packages("DAAG")
> library(DAAG)
> m3 <- lm(war_total ~ raptor_offense + raptor_defense,data=RAPTOR_CLEAN_RS)
> out <- cv.lm(data=RAPTOR_CLEAN_RS,form.lm=formula(m3),plotit="Observed",m=10)


Sum of squares = 10587    Mean square = 6.17    n = 1717

Overall (Sum over all 1717 folds)
   ms
5.95
```



Play-offs [PO] Models:

```
> RAPTOR_CLEAN_PO <- read.csv("C:/Users/IQ1006/Desktop/DSC 423/DSC 423 Group Project/RAPTOR_CL
EAN_PO.csv")
>   View(RAPTOR_CLEAN_PO)
> m4 <- lm(war_total ~ raptor_offense,data=RAPTOR_CLEAN_PO)
> summary(m4)
> m5 <- lm(war_total ~ raptor_defense,data=RAPTOR_CLEAN_PO)
> summary(m5)
> m6 <- lm(war_total ~ raptor_offense + raptor_defense,data=RAPTOR_CLEAN_PO)
> summary(m6)
> m10 <- lm(war_total ~ mp + poss + raptor_offense + pace_impact,data=RAPTOR_CLEAN_PO)
> summary(m10)
> m11 <- lm(war_total ~ mp + poss + raptor_defense + pace_impact,data=RAPTOR_CLEAN_PO)
> summary(m11)
> m12 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact, data=RAPTOR
_CLEAN_PO)
> summary(m12)
```

Model: m4

```
Call:
lm(formula = war_total ~ raptor_offense, data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1667 -0.3767 -0.2260  0.1010  6.1242

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.461148   0.009062   50.89   <2e-16 ***
raptor_offense 0.048706   0.001506   32.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7533 on 7028 degrees of freedom
Multiple R-squared:  0.1296,   Adjusted R-squared:  0.1295
F-statistic:  1047 on 1 and 7028 DF,  p-value: < 2.2e-16
```

Model: m5

```
Call:
lm(formula = war_total ~ raptor_defense, data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2980 -0.4016 -0.2471  0.1076  6.4771

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.423978   0.009353   45.33   <2e-16 ***
raptor_defense 0.048884   0.002375   20.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7842 on 7028 degrees of freedom
Multiple R-squared:  0.05687,  Adjusted R-squared:  0.05674
F-statistic: 423.8 on 1 and 7028 DF,  p-value: < 2.2e-16
```

Model: m6

```
Call:
lm(formula = war_total ~ raptor_offense + raptor_defense, data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5883 -0.3490 -0.2363  0.0965  6.0922

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.458167   0.008917   51.38   <2e-16 ***
raptor_offense 0.043932   0.001513   29.03   <2e-16 ***
raptor_defense 0.035213   0.002293   15.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7411 on 7027 degrees of freedom
Multiple R-squared:  0.1579,   Adjusted R-squared:  0.1576
F-statistic: 658.7 on 2 and 7027 DF,  p-value: < 2.2e-16
```

Model: m10

```
Call:
lm(formula = war_total ~ mp + poss + raptor_offense + pace_impact,
    data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7380 -0.1875  0.0165  0.1656  3.8103

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.1763210  0.0078977 -22.326  < 2e-16 ***
mp              0.0046500  0.0003170  14.671  < 2e-16 ***
poss           -0.0007657  0.0001636  -4.680 2.93e-06 ***
raptor_offense  0.0233672  0.0009840  23.746  < 2e-16 ***
pace_impact     0.0258942  0.0043694   5.926 3.25e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4474 on 7025 degrees of freedom
Multiple R-squared:  0.6932,   Adjusted R-squared:  0.693
F-statistic:  3968 on 4 and 7025 DF,  p-value: < 2.2e-16
```

Model: m11

```
Call:
lm(formula = war_total ~ mp + poss + raptor_defense + pace_impact,
    data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8001 -0.1953  0.0296  0.1867  3.9320

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.1957523  0.0079578 -24.599  < 2e-16 ***
mp              0.0044048  0.0003232  13.629  < 2e-16 ***
poss           -0.0006073  0.0001667  -3.642 0.000272 ***
raptor_defense  0.0236612  0.0013992  16.910  < 2e-16 ***
pace_impact    -0.0126569  0.0041963  -3.016 0.002569 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4558 on 7025 degrees of freedom
Multiple R-squared:  0.6815,   Adjusted R-squared:  0.6813
F-statistic:  3758 on 4 and 7025 DF,  p-value: < 2.2e-16
```

Model: m12

```
Call:
lm(formula = war_total ~ mp + poss + raptor_offense + raptor_defense +
    pace_impact, data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8059 -0.1809  0.0233  0.1486  3.8318

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.1680998  0.0078338 -21.458  < 2e-16 ***
mp              0.0044766  0.0003136  14.274  < 2e-16 ***
poss           -0.0006980  0.0001618  -4.313 1.63e-05 ***
raptor_offense  0.0207831  0.0009934  20.920  < 2e-16 ***
raptor_defense  0.0177764  0.0013865  12.821  < 2e-16 ***
pace_impact     0.0193465  0.0043496   4.448 8.80e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4423 on 7024 degrees of freedom
Multiple R-squared:  0.7002,   Adjusted R-squared:    0.7
```

Play-Offs step AIC

```
> install.packages("MASS")
> library(MASS)
> m12 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact,data=RAPTOR_
CLEAN_PO)
> step <- stepAIC(m12,direction="backward")
> model_empty <- lm(war_total ~ 1, data=RAPTOR_CLEAN_PO)
> step <- stepAIC(model_empty,direction="forward",scope=list(upper=m12,lower=model_empty))
```

"backward" stepAIC

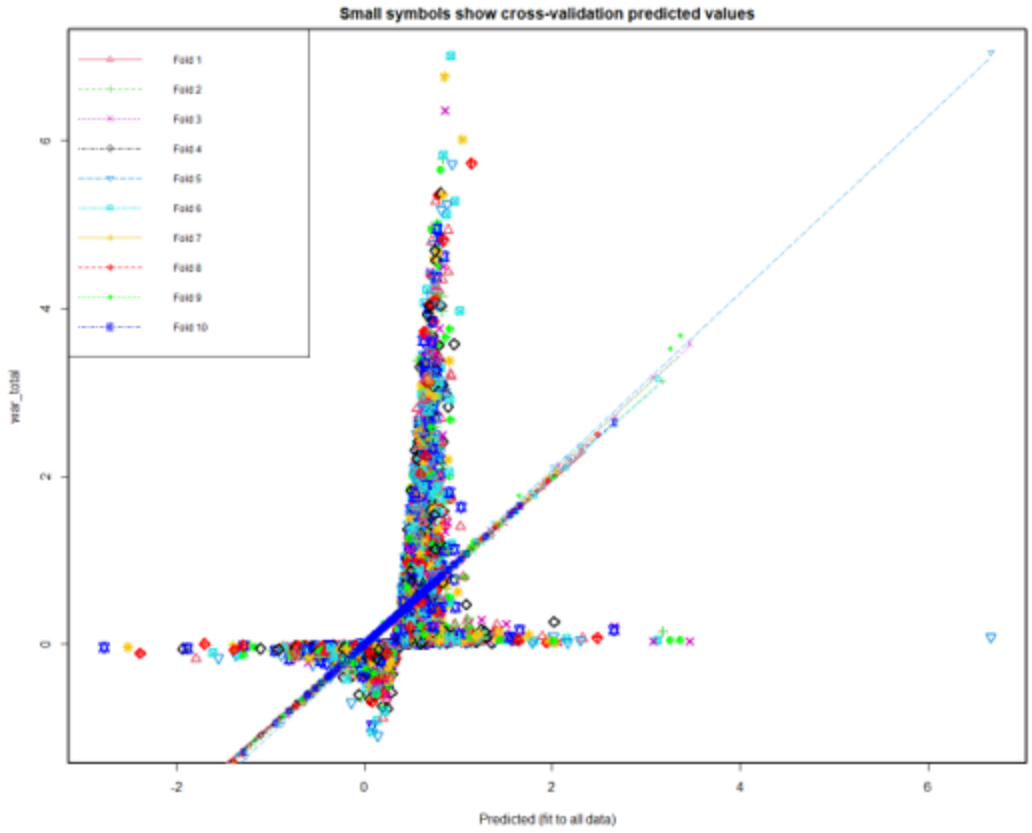
```
Start:  AIC=-11465
war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact

                 Df Sum of Sq  RSS    AIC
<none>                        1374 -11465
- poss            1       3.6 1377 -11448
- pace_impact     1       3.9 1378 -11447
- raptor_defense  1      32.2 1406 -11304
- mp              1      39.9 1414 -11266
- raptor_offense  1      85.6 1459 -11042
```

"forward" stepAIC

```
Start:  AIC=-3006
war_total ~ 1

                 Df Sum of Sq  RSS    AIC
+ mp              1      3059 1523 -10747
+ poss            1      3020 1563 -10567
+ raptor_offense  1       594 3989  -3980
+ raptor_defense  1       261 4322  -3416
+ pace_impact     1       138 4444  -3220
<none>                        4582  -3006

Step:  AIC=-10747
war_total ~ mp

                 Df Sum of Sq  RSS    AIC
+ raptor_offense  1     105.9 1417 -11252
+ raptor_defense  1      59.3 1464 -11024
+ poss            1       3.5 1520 -10761
+ pace_impact     1       1.0 1522 -10749
<none>                        1523 -10747

Step:  AIC=-11252
war_total ~ mp + raptor_offense

                 Df Sum of Sq  RSS    AIC
+ raptor_defense  1      36.1 1381 -11431
+ pace_impact     1       7.0 1410 -11285
+ poss            1       4.4 1413 -11271
<none>                        1417 -11252

Step:  AIC=-11431
war_total ~ mp + raptor_offense + raptor_defense

               Df Sum of Sq  RSS    AIC
+ pace_impact  1      3.83 1377 -11448
+ poss         1      3.60 1378 -11447
<none>                     1381 -11431

Step:  AIC=-11448
war_total ~ mp + raptor_offense + raptor_defense + pace_impact

       Df Sum of Sq  RSS    AIC
+ poss  1      3.64 1374 -11465
<none>              1377 -11448

Step:  AIC=-11465
war_total ~ mp + raptor_offense + raptor_defense + pace_impact +
    poss
```

Play-Offs N-fold Cross-Validation

```
> install.packages("DAAG")
> library(DAAG)
> m6 <- lm(war_total ~ raptor_offense + raptor_defense,data=RAPTOR_CLEAN_PO)
> out <- cv.lm(data=RAPTOR_CLEAN_PO,form.lm=formula(m6),plotit="Observed",m=10)

Sum of squares = 380    Mean square = 0.54    n = 703

Overall (Sum over all 703 folds)
   ms
0.551
```



Phase 2…

Regular Season [RS] Models:

```
> RAPTOR_CLEAN_RS <- read.csv("C:/Users/IQ1006/Desktop/DSC 423/DSC 423 Group Project/RAPTOR_CL
EAN_RS.csv")
>   View(RAPTOR_CLEAN_RS)

> RAPTOR_CLEAN_RS$raptorOD <- RAPTOR_CLEAN_RS$raptor_offense * RAPTOR_CLEAN_RS$raptor_defense
> model1 <- lm(war_total ~ raptor_offense + raptor_defense + raptorOD, data=RAPTOR_CLEAN_RS)
> summary(model1)


> RAPTOR_CLEAN_RS$raptor_defense2 <- RAPTOR_CLEAN_RS$raptor_defense * RAPTOR_CLEAN_RS$raptor_d
efense
> RAPTOR_CLEAN_RS$raptor_offense2 <- RAPTOR_CLEAN_RS$raptor_offense * RAPTOR_CLEAN_RS$raptor_o
ffense

> model2 <- lm(war_total ~ raptor_defense + raptor_defense2 + raptor_offense + raptor_offense2
,data=RAPTOR_CLEAN_RS)
> summary(model2)


> model3 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptorO
D,data=RAPTOR_CLEAN_RS)
> summary(model3)
```

Model 1:

```
Call:
lm(formula = war_total ~ raptor_offense + raptor_defense + raptorOD,
    data = RAPTOR_CLEAN_RS)

Residuals:
    Min      1Q  Median      3Q     Max
-33.224  -1.212  -0.674   0.719  38.107

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.5315607  0.0187491  135.02   <2e-16 ***
raptor_offense 0.5010448  0.0051481   97.33   <2e-16 ***
raptor_defense 0.4791466  0.0086098   55.65   <2e-16 ***
raptorOD       0.0352827  0.0006604   53.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.254 on 17169 degrees of freedom
Multiple R-squared:  0.4459,   Adjusted R-squared:  0.4458
F-statistic:  4605 on 3 and 17169 DF,  p-value: < 2.2e-16
```

Model 2:

```
Call:
lm(formula = war_total ~ raptor_defense + raptor_defense2 + raptor_offense +
    raptor_offense2, data = RAPTOR_CLEAN_RS)

Residuals:
    Min      1Q  Median      3Q     Max
-44.050  -1.187  -0.678   0.726  15.935

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.4063689  0.0190911  126.05   <2e-16 ***
raptor_defense  0.4318724  0.0088896   48.58   <2e-16 ***
raptor_defense2 0.0051456  0.0004528   11.36   <2e-16 ***
raptor_offense  0.4800526  0.0053086   90.43   <2e-16 ***
raptor_offense2 0.0085556  0.0002176   39.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.312 on 17168 degrees of freedom
Multiple R-squared:  0.4171,   Adjusted R-squared:  0.4169
F-statistic:  3071 on 4 and 17168 DF,  p-value: < 2.2e-16
```

Model 3:

```
Call:
lm(formula = war_total ~ mp + poss + raptor_offense + raptor_defense +
    pace_impact + raptorOD, data = RAPTOR_CLEAN_RS)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4975  -0.9111  -0.1944   0.6667  29.4094

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.099e-02  3.059e-02  -0.359    0.719
mp              2.410e-03  1.543e-04  15.617  < 2e-16 ***
poss           -3.180e-04  7.564e-05  -4.204 2.64e-05 ***
raptor_offense  2.783e-01  4.865e-03  57.213  < 2e-16 ***
raptor_defense  3.180e-01  7.104e-03  44.766  < 2e-16 ***
pace_impact     2.348e-01  1.701e-02  13.802  < 2e-16 ***
raptorOD        2.313e-02  5.442e-04  42.496  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.808 on 17166 degrees of freedom
Multiple R-squared:  0.6436,   Adjusted R-squared:  0.6434
F-statistic:  5166 on 6 and 17166 DF,  p-value: < 2.2e-16
```

Regular Season stepAIC:

```
> install.packages("MASS")
> library(MASS)
> step <- stepAIC(model3, direction="backward")
> model_empty <- lm(war_total ~ 1, data=RAPTOR_CLEAN_RS)
> step <- stepAIC(model_empty,direction="forward",scope=list(upper=model3,lower=model_empty))
```

"backward" stepAIC:

```
Start:  AIC=20341.7
war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact +
    raptorOD

                Df Sum of Sq   RSS   AIC
<none>                      56095 20342
- poss           1      57.8 56152 20357
- pace_impact    1     622.5 56717 20529
- mp             1     797.0 56892 20582
- raptorOD       1    5901.3 61996 22058
- raptor_defense 1    6548.7 62643 22236
- raptor_offense 1   10696.3 66791 23337
```

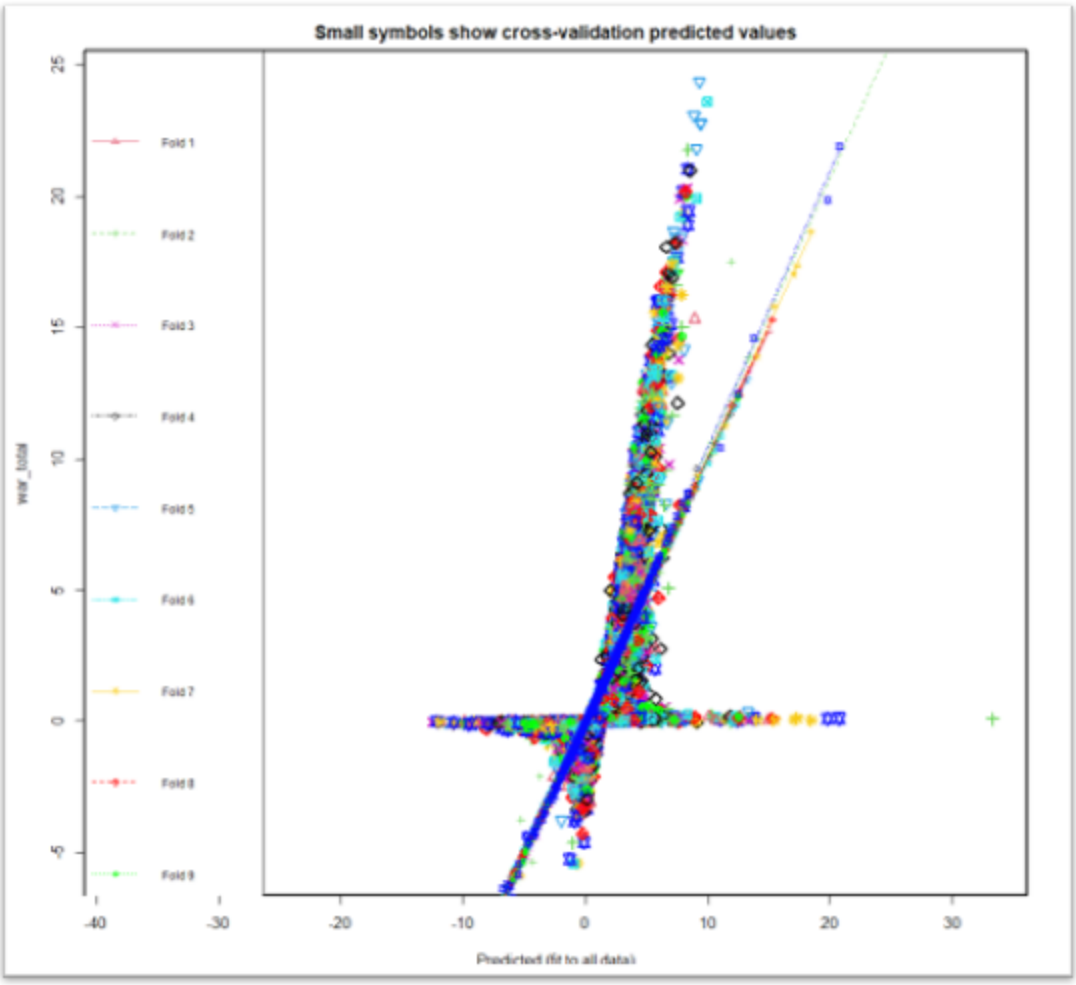
"forward" stepAIC:

```
Start:  AIC=38045.66
war_total ~ 1

                Df Sum of Sq    RSS   AIC
+ mp             1     84245  73134 24887
+ poss           1     82834  74545 25215
+ raptor_offense  1     46516 110863 32031
+ raptor_defense  1     20753 136626 35619
+ pace_impact     1     11060 146319 36796
+ raptorOD        1       263 157116 38019
<none>                      157379 38046

Step:  AIC=24886.83
war_total ~ mp

                Df Sum of Sq   RSS   AIC
+ raptor_offense  1    6711.0 66423 23236
+ raptor_defense  1    4890.7 68243 23700
+ poss           1     135.4 72998 24857
+ raptorOD        1     115.6 73018 24862
+ pace_impact     1      89.7 73044 24868
<none>                      73134 24887

Step:  AIC=23235.93
war_total ~ mp + raptor_offense

                Df Sum of Sq   RSS   AIC
+ raptor_defense  1    3611.3 62811 22278
+ raptorOD        1    2999.2 63423 22445
+ pace_impact     1     748.3 65674 23043
+ poss           1     104.9 66318 23211
<none>                      66423 23236

Step:  AIC=22277.92
war_total ~ mp + raptor_offense + raptor_defense

              Df Sum of Sq   RSS   AIC
+ raptorOD     1    6045.4 56766 20542
+ pace_impact  1     736.1 62075 22078
+ poss         1      68.2 62743 22261
<none>                    62811 22278

Step:  AIC=20542.03
war_total ~ mp + raptor_offense + raptor_defense + raptorOD

              Df Sum of Sq   RSS   AIC
+ pace_impact  1    613.66 56152 20357
+ poss         1     48.87 56717 20529
<none>                    56766 20542

Step:  AIC=20357.37
war_total ~ mp + raptor_offense + raptor_defense + raptorOD +
    pace_impact

       Df Sum of Sq   RSS   AIC
+ poss  1     57.75 56095 20342
<none>              56152 20357

Step:  AIC=20341.7
war_total ~ mp + raptor_offense + raptor_defense + raptorOD +
    pace_impact + poss
```

Regular Season n-fold Cross-Validation:

```
> install.packages("DAAG")
> library(DAAG)
> out <- cv.lm(data=RAPTOR_CLEAN_RS,form.lm=formula(model1),plotit="observed",m=10
```

```
Sum of squares = 10958    Mean square = 6.38    n = 1717

Overall (Sum over all 1717 folds)
   ms
5.15
```



Checking for Multicollinearity:

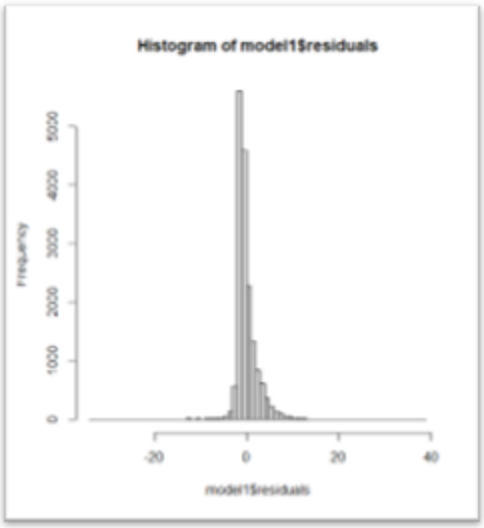
```
> install.packages("car")
> library(car)
> d <- RAPTOR_CLEAN_RS[,6:17]
> cor(d)
                    poss     mp raptor_offense raptor_defense raptor_total war_total war_reg_se
ason
poss               1.000  0.995          0.495          0.259         0.509     0.725          0
.725
mp                 0.995  1.000          0.498          0.264         0.514     0.732          0
.732
raptor_offense     0.495  0.498          1.000          0.237         0.898     0.544          0
.544
raptor_defense     0.259  0.264          0.237          1.000         0.639     0.363          0
.363
raptor_total       0.509  0.514          0.898          0.639         1.000     0.594          0
.594
war_total          0.725  0.732          0.544          0.363         0.594     1.000          1
.000
war_reg_season     0.725  0.732          0.544          0.363         0.594     1.000          1
.000
war_playoffs          NA     NA             NA             NA            NA        NA
NA
predator_offense   0.533  0.536          0.990          0.270         0.906     0.587          0
.587
predator_defense   0.355  0.360          0.323          0.929         0.675     0.422          0
.422
predator_total     0.556  0.562          0.879          0.627         0.979     0.627          0
.627
pace_impact       -0.388 -0.392         -0.364         -0.124        -0.344    -0.265         -0
.265
                  war_playoffs predator_offense predator_defense predator_total pace_impact
poss                        NA            0.533           0.3554          0.556     -0.3882
mp                          NA            0.536           0.3603          0.562     -0.3923
raptor_offense              NA            0.990           0.3227          0.879     -0.3636
raptor_defense              NA            0.270           0.9288          0.627     -0.1239
raptor_total                NA            0.906           0.6752          0.979     -0.3438
war_total                   NA            0.587           0.4220          0.627     -0.2651
war_reg_season              NA            0.587           0.4220          0.627     -0.2651
war_playoffs                 1               NA               NA             NA          NA
predator_offense            NA            1.000           0.3594          0.903     -0.3350
predator_defense            NA            0.359           1.0000          0.725     -0.0053
predator_total              NA            0.903           0.7253          1.000     -0.2496
pace_impact                 NA           -0.335          -0.0053         -0.250      1.0000
Warning message:

> vif(model1)
raptor_offense raptor_defense       raptorOD


> vif(model3)
            mp           poss raptor_offense raptor_defense    pace_impact       raptorOD
        110.82         109.68           1.75           1.22           1.24           1.43
```

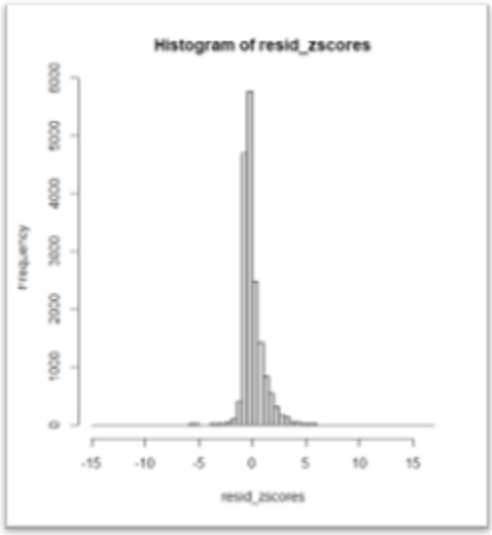
Regular Season [RS] Residual Analysis:

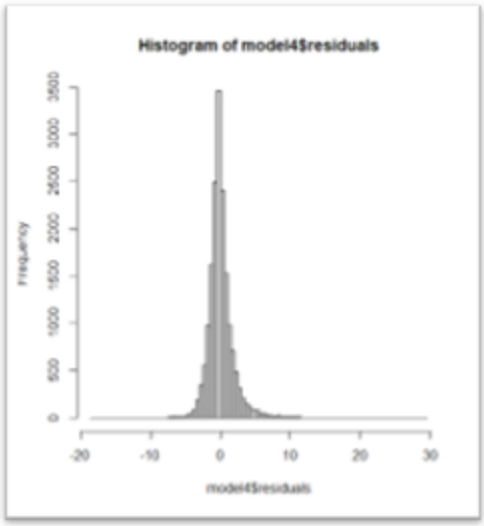
```
#Mean=0
> sum(model1$residuals)
[1] 1.04e-11
> sum(model3$residuals)
[1] 9.58e-12
#Dubrin-Watson Test

> install.packages("car")
> library(car)
> durbinWatsonTest(model1)
 lag Autocorrelation D-W Statistic p-value
   1           0.026          1.95   0.004
 Alternative hypothesis: rho != 0
> durbinWatsonTest(model3)
 lag Autocorrelation D-W Statistic p-value
   1          0.0408          1.92       0
 Alternative hypothesis: rho != 0
```

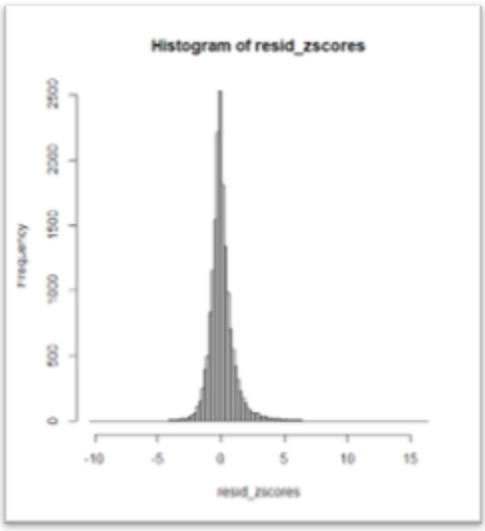
#Normal Distribution

```
> hist(model1$residuals,breaks=100)
> mean=mean(model1$residuals)
> sd=sd(model1$residuals)
> resid_zscores = (model1$residuals-mean)/sd
> hist(resid_zscores,breaks=100)

> hist(model3$residuals,breaks=100)
> mean=mean(model3$residuals)
> sd=sd(model3$residuals)
> resid_zscores=(model3$residuals-mean)/sd
> hist(resid_zscores,breaks=100)
```









Play-offs [PO] Models:

```
> RAPTOR_CLEAN_PO <- read.csv("C:/Users/IQ1006/Desktop/DSC 423/DSC 423 Group Project/RAPTOR_CL
EAN_PO.csv")
>    View(RAPTOR_CLEAN_PO)

> RAPTOR_CLEAN_RS$raptorOD <- RAPTOR_CLEAN_RS$raptor_offense * RAPTOR_CLEAN_RS$raptor_defense
> model4 <- lm(war_total ~ raptor_offense + raptor_defense + raptorOD, data=RAPTOR_CLEAN_PO)
> summary(model4)

> RAPTOR_CLEAN_PO$raptor_defense2 <- RAPTOR_CLEAN_PO$raptor_defense * RAPTOR_CLEAN_PO$raptor_d
efense
> RAPTOR_CLEAN_PO$raptor_offense2 <- RAPTOR_CLEAN_PO$raptor_offense * RAPTOR_CLEAN_PO$raptor_o
ffense
> model5 <- lm(war_total ~ raptor_defense + raptor_defense2 + raptor_offense + raptor_offense2
,data=RAPTOR_CLEAN_PO)


> model6 <- lm(war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact + raptor_
defense2 + raptor_offense2,data=RAPTOR_CLEAN_PO)
> summary(model6)
```

Model 4:

```
Call:
lm(formula = war_total ~ raptor_offense + raptor_defense + raptorOD,
    data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3296 -0.3497 -0.2379  0.0968  6.0958

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.4599759  0.0089219  51.556  < 2e-16 ***
raptor_offense  0.0438781  0.0015120  29.020  < 2e-16 ***
raptor_defense  0.0362026  0.0023062  15.698  < 2e-16 ***
raptorOD       -0.0003778  0.0001018  -3.713 0.000207 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7404 on 7026 degrees of freedom
Multiple R-squared:  0.1595,   Adjusted R-squared:  0.1592
F-statistic: 444.5 on 3 and 7026 DF,  p-value: < 2.2e-16
```

Model 5:

```
Call:
lm(formula = war_total ~ raptor_defense + raptor_defense2 + raptor_offense +
    raptor_offense2, data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7023 -0.3539 -0.2425  0.0970  6.0635

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.757e-01  9.004e-03  52.835   <2e-16 ***
raptor_defense  5.086e-02  2.588e-03  19.654   <2e-16 ***
raptor_defense2 -8.346e-04  6.776e-05 -12.317   <2e-16 ***
raptor_offense  4.296e-02  1.494e-03  28.747   <2e-16 ***
raptor_offense2 -1.386e-04  5.525e-05  -2.509   0.0121 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.731 on 7025 degrees of freedom
Multiple R-squared:  0.1808,   Adjusted R-squared:  0.1803
F-statistic: 387.5 on 4 and 7025 DF,  p-value: < 2.2e-16
```

Model 6:

```
Call:
lm(formula = war_total ~ mp + poss + raptor_offense + raptor_defense +
    pace_impact + raptor_defense2 + raptor_offense2, data = RAPTOR_CLEAN_PO)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9500 -0.1774  0.0232  0.1512  3.8008

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.761e-01  8.075e-03 -21.812  < 2e-16 ***
mp               4.433e-03  3.113e-04  14.239  < 2e-16 ***
poss            -6.720e-04  1.606e-04  -4.183 2.91e-05 ***
raptor_offense   2.106e-02  9.866e-04  21.349  < 2e-16 ***
raptor_defense   2.360e-02  1.574e-03  15.047  < 2e-16 ***
pace_impact      2.547e-02  4.362e-03   5.838 5.52e-09 ***
raptor_defense2 -3.361e-04  4.136e-05  -8.126 5.18e-16 ***
raptor_offense2  2.925e-04  3.350e-05   8.731  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4389 on 7022 degrees of freedom
Multiple R-squared:  0.7048,   Adjusted R-squared:  0.7045
F-statistic:  2395 on 7 and 7022 DF,  p-value: < 2.2e-16
```

Play-offs Season stepAIC:

```
> install.packages("MASS")
> library(MASS)
> step <- stepAIC(model6,direction="backward")

> model_empty <- lm(war_total ~ 1, data=RAPTOR_CLEAN_PO)
> step <- stepAIC(model_empty,direction="forward",scope=list(upper=model6,lower=model_empty))
```

"backward" stepAIC:

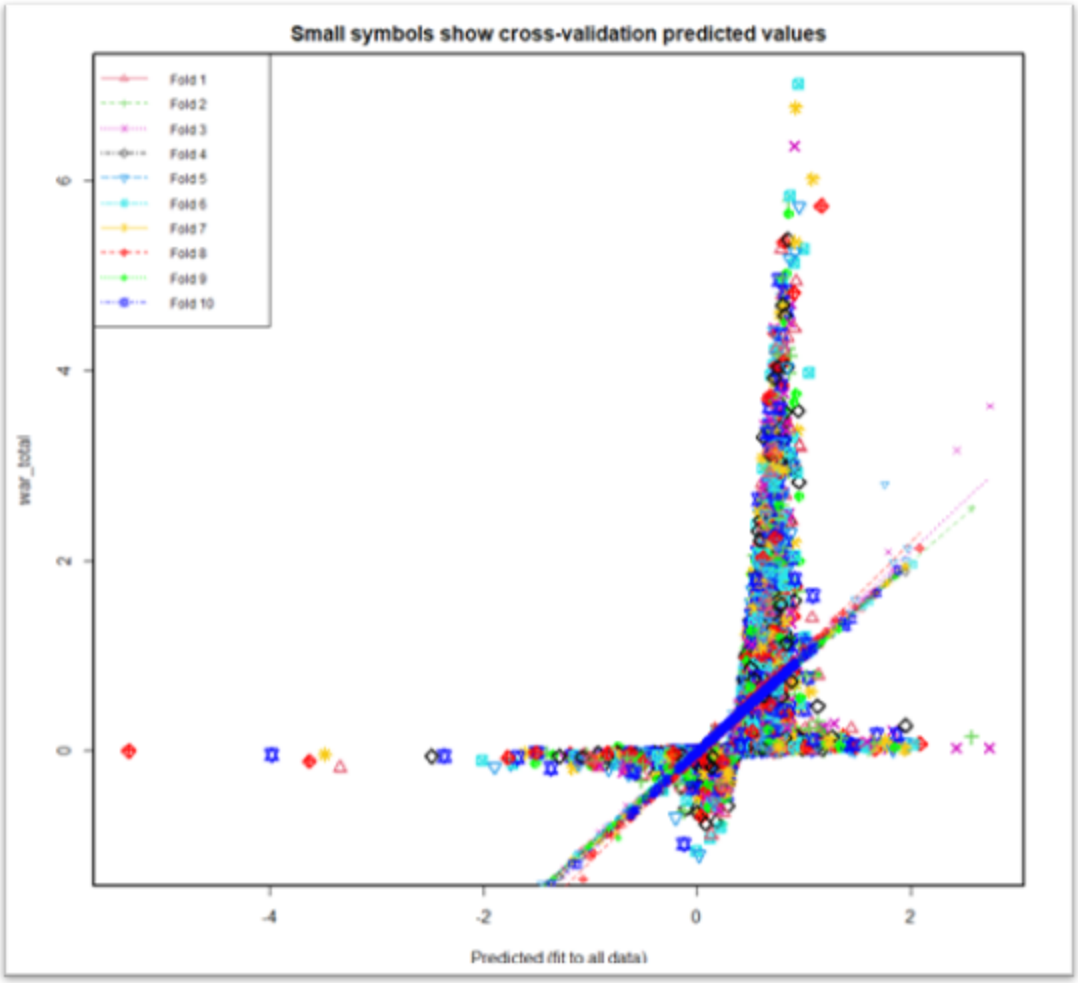
```
Start:  AIC=-11569.72
war_total ~ mp + poss + raptor_offense + raptor_defense + pace_impact +
    raptor_defense2 + raptor_offense2

                  Df Sum of Sq    RSS    AIC
<none>                        1352.8 -11570
- poss             1     3.371 1356.1 -11554
- pace_impact      1     6.566 1359.3 -11538
- raptor_defense2  1    12.722 1365.5 -11506
- raptor_offense2  1    14.686 1367.5 -11496
- mp               1    39.060 1391.8 -11372
- raptor_defense   1    43.620 1396.4 -11349
- raptor_offense   1    87.804 1440.6 -11130
```

"forward" stepAIC:

```
Start:  AIC=-3006.45
war_total ~ 1

                  Df Sum of Sq    RSS      AIC
+ mp               1   3059.17 1523.3 -10747.0
+ poss             1   3019.75 1562.7 -10567.3
+ raptor_offense   1    593.93 3988.6  -3980.3
+ raptor_defense   1    260.63 4321.9  -3416.1
+ pace_impact      1    138.47 4444.0  -3220.2
+ raptor_offense2  1     26.96 4555.5  -3045.9
+ raptor_defense2  1      4.68 4577.8  -3011.6
<none>                        4582.5  -3006.5

Step:  AIC=-10746.95
war_total ~ mp

                  Df Sum of Sq    RSS    AIC
+ raptor_offense   1   105.932 1417.4 -11252
+ raptor_defense   1    59.253 1464.1 -11024
+ raptor_offense2  1    10.344 1513.0 -10793
+ poss             1     3.494 1519.8 -10761
+ raptor_defense2  1     2.867 1520.5 -10758
+ pace_impact      1     0.956 1522.4 -10749
<none>                        1523.3 -10747

Step:  AIC=-11251.65
war_total ~ mp + raptor_offense

                  Df Sum of Sq    RSS    AIC
+ raptor_defense   1    36.079 1381.3 -11431
+ raptor_offense2  1     8.913 1408.5 -11294
+ pace_impact      1     7.013 1410.4 -11284
+ poss             1     4.367 1413.0 -11271
+ raptor_defense2  1     0.800 1416.6 -11254
<none>                        1417.4 -11252

Step:  AIC=-11430.91
war_total ~ mp + raptor_offense + raptor_defense

                  Df Sum of Sq    RSS    AIC
+ raptor_offense2  1    7.9599 1373.3 -11470
+ raptor_defense2  1    5.2629 1376.0 -11456
+ pace_impact      1    3.8304 1377.5 -11448
+ poss             1    3.5990 1377.7 -11447
<none>                        1381.3 -11431

Step:  AIC=-11469.54
war_total ~ mp + raptor_offense + raptor_defense + raptor_offense2

                  Df Sum of Sq    RSS    AIC
+ raptor_defense2  1   10.6800 1362.7 -11522
+ pace_impact      1    4.3400 1369.0 -11490
+ poss             1    3.4874 1369.9 -11485
<none>                        1373.3 -11470

Step:  AIC=-11522.43
war_total ~ mp + raptor_offense + raptor_defense + raptor_offense2 +
    raptor_defense2

              Df Sum of Sq    RSS    AIC
+ pace_impact  1    6.5362 1356.1 -11554
+ poss         1    3.3411 1359.3 -11538
<none>                     1362.7 -11522

Step:  AIC=-11554.23
war_total ~ mp + raptor_offense + raptor_defense + raptor_offense2 +
    raptor_defense2 + pace_impact

       Df Sum of Sq    RSS    AIC
+ poss  1    3.3707 1352.8 -11570
<none>              1356.1 -11554

Step:  AIC=-11569.72
war_total ~ mp + raptor_offense + raptor_defense + raptor_offense2 +
    raptor_defense2 + pace_impact + poss
```

Play-offs n-fold Cross-Validation:

```
> install.packages("DAAG")
> library(DAAG)
> out <- cv.lm(data=RAPTOR_CLEAN_PO,form.lm=formula(models),plotit="observed",m=10

Sum of squares = 386    Mean square = 0.55    n = 703

Overall (Sum over all 703 folds)
   ms
0.54
```



Checking for Multicollinearity:

```
> install.packages("car")
> library(car)
> d <- RAPTOR_CLEAN_PO[,6:17]
> cor(d)
                    poss     mp raptor_offense raptor_defense raptor_total war_total war_reg_se
ason
poss               1.000  0.996          0.262         0.1515        0.277     0.812
NA
mp                 0.996  1.000          0.261         0.1544        0.278     0.817
NA
raptor_offense     0.262  0.261          1.000         0.2054        0.869     0.360
NA
raptor_defense     0.151  0.154          0.205         1.0000        0.662     0.238
NA
raptor_total       0.277  0.278          0.869         0.6624        1.000     0.396
NA
war_total          0.812  0.817          0.360         0.2385        0.396     1.000
NA
war_reg_season        NA     NA             NA             NA           NA        NA
1
war_playoffs       0.812  0.817          0.360         0.2385        0.396     1.000
NA
predator_offense   0.295  0.294          0.988         0.2585        0.887     0.391
NA
predator_defense   0.210  0.213          0.256         0.9494        0.676     0.275
NA
predator_total     0.319  0.319          0.841         0.6671        0.981     0.421
NA
pace_impact       -0.195 -0.195         -0.370         0.0187       -0.274    -0.174
NA
                  war_playoffs predator_offense predator_defense predator_total pace_impact
poss                     0.812           0.295           0.210          0.319     -0.1952
mp                       0.817           0.294           0.213          0.319     -0.1954
raptor_offense           0.360           0.988           0.256          0.841     -0.3703
raptor_defense           0.238           0.258           0.949          0.667      0.0187
raptor_total             0.396           0.887           0.676          0.981     -0.2740
war_total                1.000           0.391           0.275          0.421     -0.1738
war_reg_season              NA              NA              NA             NA         NA
war_playoffs             1.000           0.391           0.275          0.421     -0.1738
predator_offense         0.391           1.000           0.309          0.876     -0.3444
predator_defense         0.275           0.309           1.000          0.729      0.1421
predator_total           0.421           0.876           0.729          1.000     -0.1758
pace_impact             -0.174          -0.344           0.142         -0.176      1.0000


> vif(model5)
 raptor_defense raptor_defense2  raptor_offense raptor_offense2
           1.37            1.42            1.05            1.11
> vif(model6)
             mp            poss  raptor_offense  raptor_defense     pace_impact raptor_defense
2
         139.51          139.42            1.26            1.40            1.21             1.4
7
raptor_offense2
           1.13
```

Play-offs [PO] Residual Analysis:

#Mean=0

```
> sum(model5$residuals)

[1] 1.11e-13
> sum(model6$residuals)
[1] 5.08e-14
```

#Durbin-Watson Test

```
> install.packages("car")
> library(car)
> durbinWatsonTest(model5)
 lag Autocorrelation D-W Statistic p-value
   1           0.144          1.71       0
 Alternative hypothesis: rho != 0
> durbinWatsonTest(model6)
 lag Autocorrelation D-W Statistic p-value
   1         -0.0493          2.1       0
 Alternative hypothesis: rho != 0
```

#Normal Distribution

```
> hist(model5$residuals,breaks=100)
> mean = mean(model5$residuals)
> sd=sd(model5$residuals)
> resid_zscores = (model5$residuals-mean)/sd
> hist(resid_zscores,breaks=100)

> hist(model6$residuals,breaks=100)
> mean=mean(model6$residuals)
> sd=sd(model6$residuals)
> resid_zscores = (model6$residuals - mean)/sd
> hist(resid_zscores,breaks=100)
```