Report of factors on basketball players' performance

Introduction

The use of data analysis is attracting more attention from sports institutions. While countries and sports teams spend a lot of money on efforts in winning games, data analysis gives them reference on maximizing the efficiency with the costs(Sarlis & Tjortjis, 2020). In order to limit the gap of the wealth of teams' backgrounds, NBA punishes teams that exceed the salary cap by reducing their privileges in free agency and charging them with extra taxes(Wikipedia,2021). Also, the performance of players in each team directly shows whether they are worth such salaries. Therefore, by predicting players' performance, teams are able to increase their benefit from the investment in players. For example, the research by Pehar et al(2017) provides a model that uses a jump test for basketball players in a different position to predict their performance. Therefore, we aim to find a model to find out what are factors impact the performance of basketball players from other approaches in order to provide references for NBA basketball teams on their activities in free agency and trading.

Method

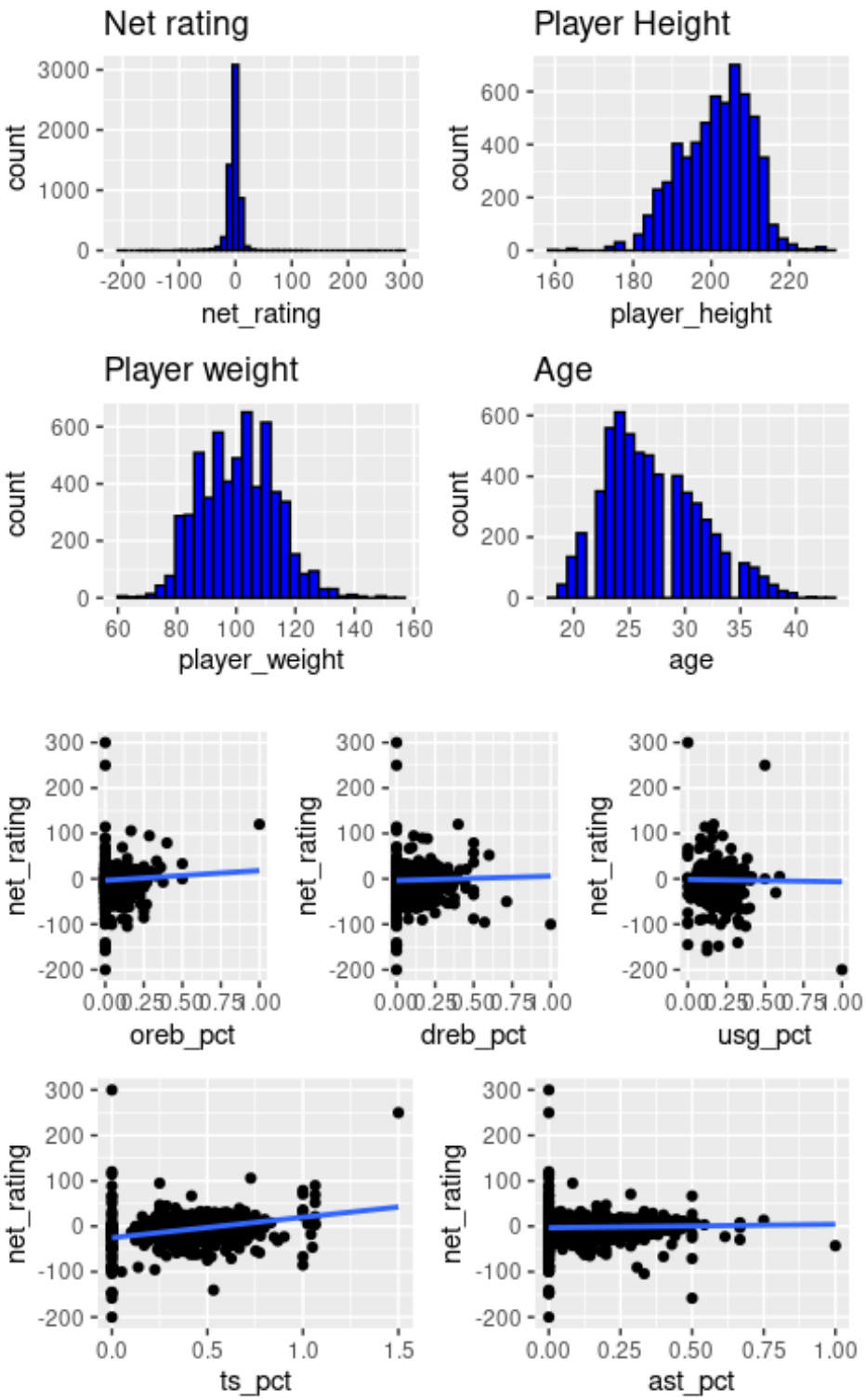
The main method we will explore is multiple linear regression (MLP). The MLP only works well when all the assumptions are held. Otherwise, the estimates and further analysis of the model will not be reliable. The description of important variables in raw data is shown in table 1 in the appendix. Our dataset is from Kaggle. It included data corresponding to demographic variables and basic box score statistics of NBA players from season 1996-97 to season 2020-21. The dataset is updated on 02/08/2021 (Justinas, 2021). Since we are going to find out what are factors influence player's performance from the way they can benefit the team, the most appropriate response variable should be net rating which indicates the team's point differential per 100 possessions while the certain player is on the court (Justinas, 2021). Furthermore, since we plan to generate a validation for our model, we first split our dataset into train data and test. Thus, we will use only train data until the validation section. The general description of variables is in table 1 of appendix.

The structure of the Method selection follows:

- Exploratory Data Analysis (EDA)
- Starting model analyze
- Model comparison
- Model validation

EDA

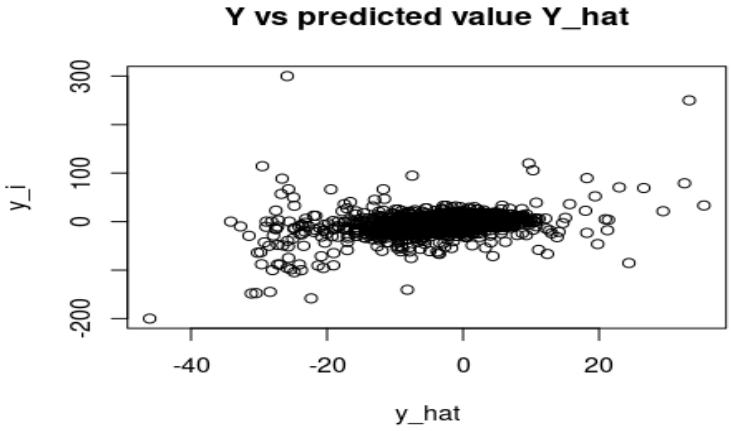
Figure 1


Net rating
count
net_rating
Player Height
count
player_height
Player weight
count
player_weight
Age
count
age
net_rating
oreb_pct
net_rating
dreb_pct
net_rating
usg_pct
net_rating
ts_pct
net_rating
ast_pct

Referring to figure 1, we can see the body data associated with players are various. Also, there is no observation that is shown unexpectedly. The net rating is highly concentrated around 0. With regards to scatterplots, we selected 5 important performance-related variables and see the relationship with our response variable net rating. The results show that there is some linear relationship between performance-related variables and net rating although the relationship seems not strong.

Starting model analyze

Figure 2



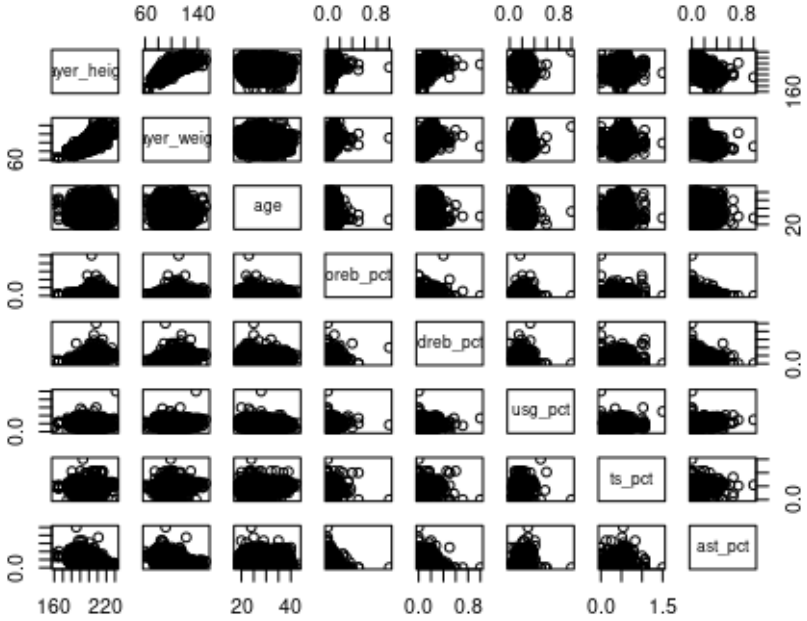
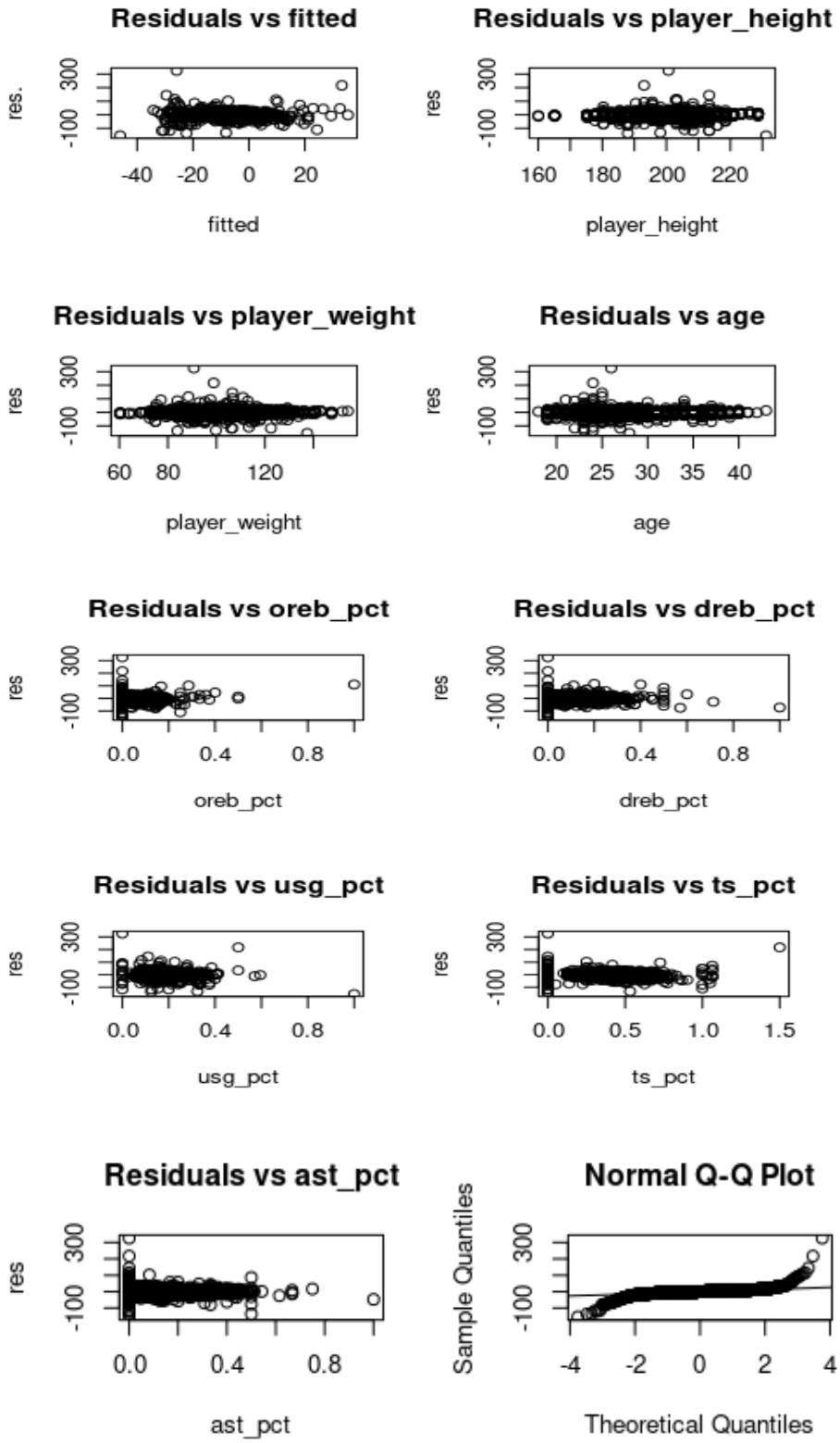


Figure 3


Residuals vs fitted
res.
fitted
Residuals vs player_height
res
player_height
Residuals vs player_weight
res
player_weight
Residuals vs age
res
age
Residuals vs oreb_pct
res
oreb_pct
Residuals vs dreb_pct
res
dreb_pct
Residuals vs usg_pct
res
usg_pct
Residuals vs ts_pct
res
ts_pct
Residuals vs ast_pct
res
ast_pct
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

As Pehar et al. (2017) pointed out, the jumping ability might affect basketball player's performance. Therefore, we add body data to the model while the jumping ability seems to be variance among players with different physical conditions. Moreover, the performance-related data might affect the team's point differential on the court. Teams with more offense rebounds typically have more opportunities to score. Therefore, we start with the model (model 1):

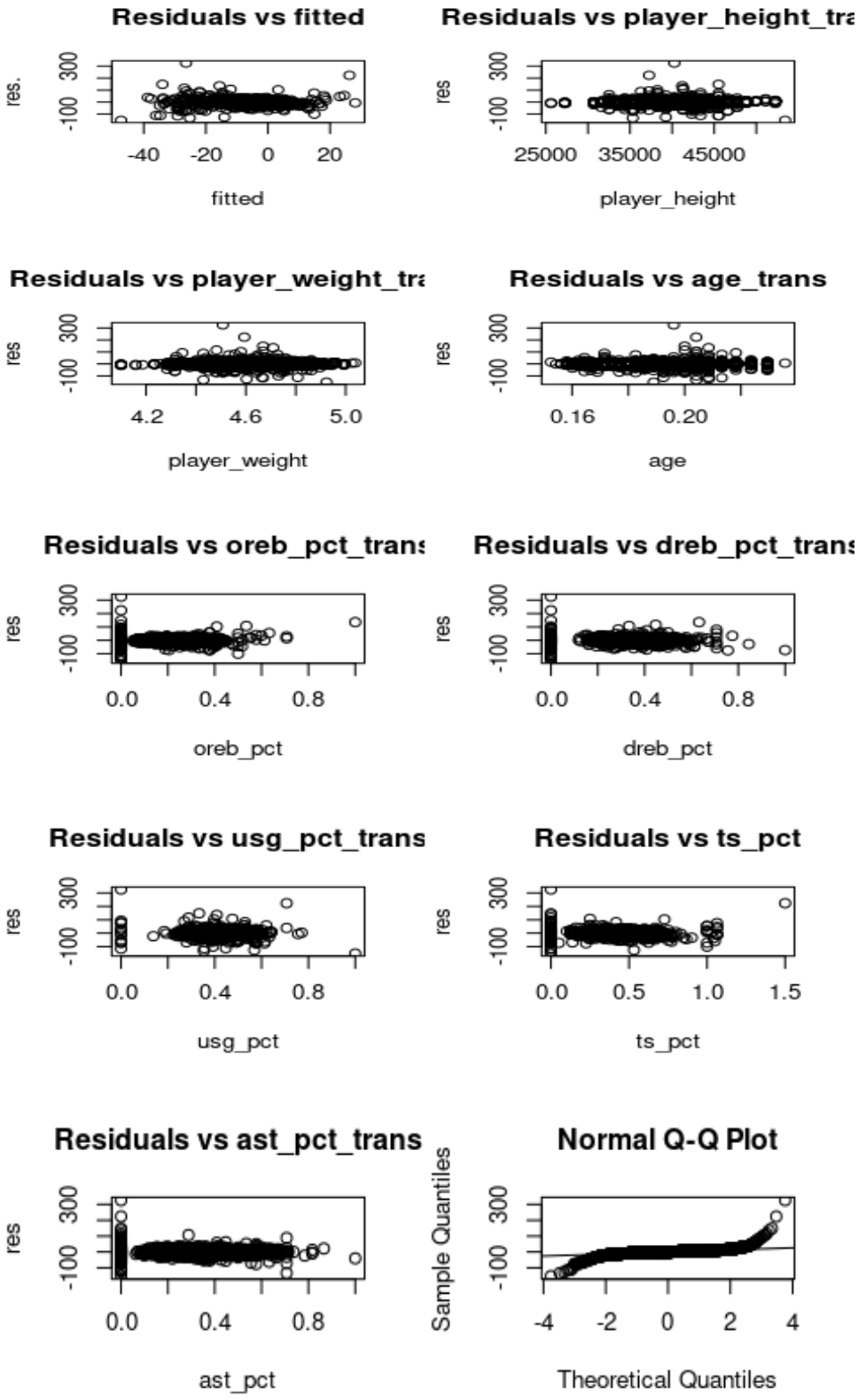
$$net\_rating \sim player\_height + player\_weight + age + oreb\_pct + dreb\_pct + usg\_pct + ts_pct + ast\_pct$$

Also, we believe GP, reb, and ast are highly collinear with other performance-related data in our model. Thus, we choose not to include them in model 1.

We need to check multicollinearity in our model and whether assumptions are held in this model. We begin with checking the multicollinearity through the variance inflation factor (VIF) of variables. The multicollinearity can cause wrong estimates of coefficients and might produce a large variance. Fortunately, in model 1, none of the variables has VIF greater than 5 which means there is no strong multicollinearity.

Then, we check the assumptions through residual plots and Q-Q plots. The residual plots allow us to verify assumptions such as linearity, normality, constant variance, and uncorrelatedness while the Q-Q plot is used to check normality. In addition, before checking the assumption, we must make sure that our model is under two conditions. Otherwise, the results of residual plots are not able to show the true issues. Referring to figure 2, we can see a clear pattern between y_i and predicted value $\hat{y}$ which satisfied the first condition. Also, there is no relationship other than the linear relationship between predictors which means condition 2 is met. Based on figure 3, we can see a few triangle shapes in a few plots. Thus, constant variance tends to be violated in our model. More importantly, there is a huge deviation from the straight diagonal string on two tails. So, the normality is also violated which means inference on the estimates is not reliable in our model. In consequence, we need to apply a transformation to the model and see if we can fix the problem.

Figure 4


Residuals vs fitted
res.
fitted
Residuals vs player_height_tra
res
player_height
Residuals vs player_weight_tra
res
player_weight
Residuals vs age_trans
res
age
Residuals vs oreb_pct_trans
res
oreb_pct
Residuals vs dreb_pct_trans
res
dreb_pct
Residuals vs usg_pct_trans
res
usg_pct
Residuals vs ts_pct
res
ts_pct
Residuals vs ast_pct_trans
res
ast_pct
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

While we use power transformation, we can only apply transformation on predictors while our respond variable has negative elements which means we are not able meet constant variance and normality. Overall, our model becomes(model 2):

$$net\_rating \sim player\_height^2 + ln(player\_weight) + \frac{1}{\sqrt{age}} + \sqrt{oreb\_pct} + \sqrt{dreb\_pct} + \sqrt{usg\_pct} + ts_pct + \sqrt{ast\_pct}$$

Overall, as shown in figure 3, the violation of normality is not fixed as expected. More importantly, the clusters in the residual plots are more clear after transformation.

Model comparison

While our model cannot meet all the assumptions and the standard error is high. We generate a model with more variables to compare(model 3):

$$net\_rating \sim gp + pts + reb + ast + player\_height + player\_weight + age + oreb\_pct + dreb\_pct + usg\_pct + ts_pct + ast\_pct$$

Surprisingly, we do not have any multicollinearity problem in model 3 with all the performance-related data which is against our hypothesis on model 1. We also apply the same process as train data to the model. However, we still have the same issues as model 1. Again, we are unable to fix the issues in normality by power transformation. Therefore, we will only validate three models in this report.

Model validation

In order to find out whether our models are overfitting the train data, we use a test dataset to validate our models by comparing their coefficients and see if new violations of assumptions are made in test data. Basically, we apply the same transformation to the variable and generate models with the same predictors as train data.

All analysis for this report was programmed using R version 4.0.2 with R package Tidyverse (Wickham et al., 2021), Huxtable (Hugh-Jones, 2021), car (John et al., 2021), and patchwork (Pedersen, 2020).

Result

Figure 5

	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	-16.674 **	-22.437 ***	16.793	-9.438	-115.181	-24.579
	(5.932)	(4.950)	(9.995)	(8.028)	(68.860)	(58.065)
player_height	-0.042	-0.031			1.485 *	0.196
	(0.035)	(0.029)			(0.733)	(0.624)
player_weight	-0.070 **	0.012			0.029	-0.017
	(0.024)	(0.020)			(0.194)	(0.163)
age	0.217 ***	0.261 ***			0.076	0.006
	(0.039)	(0.031)			(0.271)	(0.219)
oreb_pct	38.886 ***	14.645 ***			152.273 ***	47.006 ***
	(4.839)	(4.232)			(12.465)	(12.288)
dreb_pct	3.742	1.783			-84.770 ***	-43.208 ***
	(3.480)	(2.895)			(11.249)	(10.064)
usg_pct	-16.106 ***	-4.420			39.178 *	-4.193
	(3.308)	(2.759)			(17.185)	(15.281)
ts_pct	45.233 ***	32.712 ***	43.391 ***	31.067 ***	35.554 ***	20.908 ***
	(1.689)	(1.349)	(1.721)	(1.377)	(1.883)	(1.569)
ast_pct	9.915 ***	9.653 ***			-28.115 **	-22.249 **
	(2.367)	(1.964)			(8.552)	(7.124)
player_height_trans			-0.000	-0.000	-0.004 *	-0.001
			(0.000)	(0.000)	(0.002)	(0.002)
player_weight_trans			-5.675 *	1.402	-10.327	1.670
			(2.478)	(2.014)	(19.985)	(16.667)
age_trans			-57.732 ***	-74.602 ***	-27.068	-66.230
			(11.195)	(8.894)	(77.990)	(63.280)
oreb_pct_trans			13.262 ***	9.107 ***	-53.872 ***	-9.370
			(2.383)	(1.964)	(6.187)	(5.662)
dreb_pct_trans			10.191 ***	4.111 *	64.837 ***	32.528 ***
			(2.443)	(2.050)	(7.543)	(6.757)
usg_pct_trans			-16.944 ***	-6.165 **	-62.868 ***	-19.847
			(2.800)	(2.241)	(13.791)	(11.499)
ast_pct_trans			9.105 ***	9.645 ***	22.321 ***	16.053 ***
			(1.685)	(1.370)	(5.273)	(4.240)
gp					0.033 ***	0.039 ***
					(0.009)	(0.007)
pts					0.276 ***	0.337 ***
					(0.075)	(0.067)
reb					-0.022	-0.138
					(0.147)	(0.125)
ast					0.395	0.391
					(0.232)	(0.200)
N	5850	5850	5850	5850	5850	5850
R2	0.132	0.114	0.134	0.122	0.179	0.164
logLik	-23050.007	-21783.115	-23042.240	-21756.906	-22884.230	-21613.377
AIC	46120.013	43586.231	46104.480	43533.812	45810.460	43268.754

*** p < 0.001; ** p < 0.01; * p < 0.05.

The results from the models are performed in figure 5. Every two rows show the results of model 1 on train data and test data. The estimates for coefficients are close in model 1 while there is a huge gap in model 2 and model 3. Although the assumption violation is the main issue among models, there are no additional violations invalidation. From the perspective of the explanatory power of models, model 3 has the highest R^2 which means the true variance explained by model 3 is the most. With regards to AIC, we typically preferred a model with low AIC which means model 3 is also the best choice from the perspective of AIC. Overall, we tend to take model 1 as our final model because there are no significant differences in R^2 and AIC between models while the differences in estimates of model 2 and model 3 between train data and test data are significantly large.

Disscussion

We still can generate some important results. The estimates for coefficients of player_weight, age, oreb_pct, usg_pct,ts_pct, and ast_pct are statistically significant which means these are the factors that are more possible to have an impact on net_rating. Also, some results are reasonable. For example, the estimates for usg_pct mean with one unit increase in the percentage of the player using the ball the predicted net rating of the player will decrease by 16.106. It is possible because the efficiency of using the ball might increase with decreasing in the usage of the ball which decreases the performance.

However, the limitation significantly affects the reliability of this report. One is that the number of influential points is significantly huge among all models in both train and test data. Therefore, it might cause differences in estimates in train and test data. Therefore, the results of validation might not be precise. In addition, the problems might not be caused by data selection because the distribution of test and train data is similar (table 2, appendix).

Also, even though we choose model 1 as the final model, the estimates might also be biased. First, none of the violations in the assumption is fixed by the power transformation. Therefore, any results of model are not reliable. Also, R^2 is only 0.132 in final model. Therefore, the number of variances that could be explained by our model is limited. One possible reason of such issues is that our data includes all the players played in the NBA which means those who only played a few games can also appear in the sample. More importantly, while only top players can play in the league for a long period of time, the number of these observations tends to be large. Consequently, our sample data is not credible in doing such research. To fix this problem, we must find a way to clear out these players and do research only on top players. For example, Sarlis & Tjortjis (2020) only analyze data on top players.

Reference

Fox, J., Weisberg, S., and Price, Brad.(2021). car: Companion to Applied Regression, https://CRAN.R-project.org/package=car

Huge-Jones, D., (2021). huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats, https://hughjonesd.github.io/huxtable/

Justinas, Cirtautas.(2021). NBA Players: Biometric, biographic and basic box score features from 1996 to 2019 season. *Kaggle* Retrieved From https://www.kaggle.com/justinas/nba-players-data

Pehar, M., Sekulic, D., Sisic, N., Spasic, M., Uljevic, O., & Krolo, A. et al. (2017). Evaluation of different jumping tests in defining position-specific and performance-level differences in high level basketball players. *Biology of Sport*, 34(3), 263-272. https://doi.org/10.5114/biolsport.2017.67122

Perdersen Lin, T.(2020). patchwork: The Composer of Plots, https://patchwork.data-imaginist.com

Sarlis, V., & Tjortjis, C. (2020). Sports analytics — evaluation of basketball players and Team Performance. *Information Systems*, 93, 101562. https://doi.org/10.1016/j.is.2020.101562

Wikimedia Foundation. (2021, December 7). NBA salary cap. *Wikipedia*. Retrieved December 7, 2021, from https://en.wikipedia.org/wiki/NBA_salary_cap.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendix

Table 1

Variables	Description	Type
net_rating	Point differences while the player is on the court(OTC) per 100 possessions	Response
gp	game played per season	Performance-related
pts	Score points per game	Performance-related
reb	Rebound per game	Performance-related
ast	assistant per game	Performance-related
player_height	height of the player	Body data
player_weight	Weight of the player	Body data
age	age of the player	Body data
oreb_pct	offensive rebound rate while the player is OTC	Performance-related
dreb_pct	defensive rebound rate while the player is OTC	Performance-related
usg_pct	the usage of ball by the player while he is OTC	Performance-related
ts_pct	shooting efficiency of the player while he is OTC	Performance-related
ast_pct	assisted goal rate while the player is OTC	Performance-related

Table 2

Variable	mean (s.d.) in training	mean (s.d.) in test
net_rating	-2.192 (13.354)	-2.141 (10.649)
gp	51.677 (25.056)	51.757 (24.916)
pts	8.215 (6.033)	8.124 (5.878)
reb	3.566 (2.468)	3.564 (2.507)
ast	1.823 (1.823)	1.799 (1.761)
player_height	200.767 (9.26)	200.69 (9.08)
player_weight	100.646 (12.651)	100.408 (12.401)
age	27.11 (4.316)	27.154 (4.364)
oreb_pct	0.055 (0.045)	0.055 (0.042)
dreb_pct	0.141 (0.063)	0.142 (0.062)
usg_pct	0.186 (0.054)	0.185 (0.052)
ts_pct	0.511 (0.098)	0.509 (0.099)
ast_pct	0.132 (0.095)	0.131 (0.093)