Business Report

NBA League Salary

QBUS2820 - 500080960

Table of contents

Part A + Part B: 15 pages

PART A

1. Business context

The report aims to find a model that can predict salary of NBA players well. The success of the finding would provide a good benchmark for basketball teams to estimate the appropriate salary for their players, especially superstars. Some main considerations in this analysis are:

- What attributes of players can help explain their high/low salary?
- What are the magnitudes of effects of those attributes on salary?
- What model is appropriate to be used here?

It is also hypothesised that players who overall contribute effectively to the team's wins and are assigned lots of time to play should have a greater income. These questions and hypothesis would be answered through the report when three different models, which are multiple linear regression, KNN regression, and kernel regression, are fitted to choose the most appropriate one based on the NBA league's aim: predictability. RMSE test set, which shows the performance of models on new unseen data, would be the main basis for comparison.

2. Data processing

Algorithms have been run to identify missing values, and fortunately all variables have the same number of data points. Furthermore, all the data have clear headings and units without probably abnormal values observed (e.g., extremely large values). Therefore, no further data cleaning is performed. Otherwise, unreasonable manipulation would distort the true association between random variables, and this would lead to biases in interpretation.

3. Exploratory data analysis

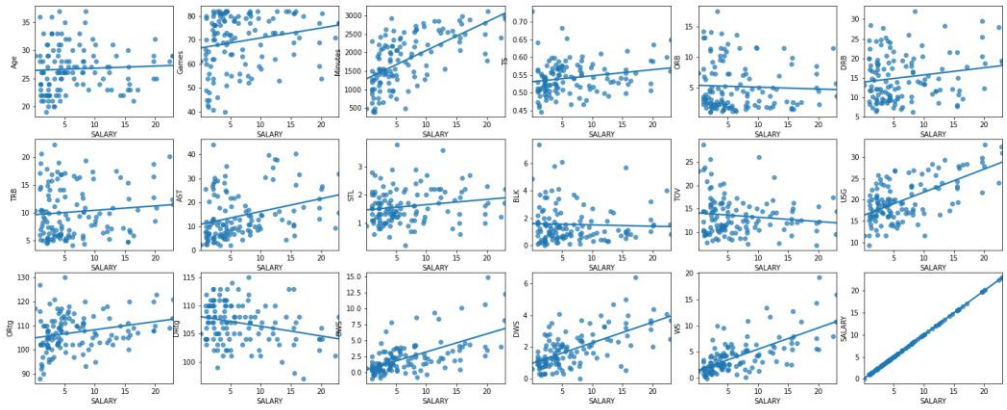

Figure 3.1. Regression plots between SALARY and potential predictors

	SALARY	Age	Games	Minutes	PER	TS	ORB	DRB	TRB	AST	STL	BLK	TOV	USG	ORtg	DRtg	OWS	DWS	WS
SALARY	1	0.0537	0.2051	0.6213	0.6702	0.2102	-0.0449	0.1735	0.0974	0.3101	0.1717	-0.0405	-0.1256	0.6196	0.2519	-0.2657	0.5982	0.6262	0.6778
Age	0.0537	1	-0.1308	-0.0114	0.1031	0.2304	-0.0794	0.0518	0.0054	0.0852	-0.0073	-0.0346	0.0695	0.0323	0.1677	-0.1174	0.0449	0.0773	0.0606
Games	0.2051	-0.1308	1	0.6986	0.2187	0.2223	-0.0383	-0.0693	-0.0605	0.0538	-0.0109	-0.0969	-0.1707	0.1380	0.2831	-0.1090	0.3878	0.4890	0.4704
Minutes	0.6213	-0.0114	0.6986	1	0.5429	0.2371	-0.2827	-0.1321	-0.2042	0.4010	0.1951	-0.2349	-0.2201	0.4952	0.3447	-0.0222	0.6612	0.6676	0.7427
PER	0.6702	0.1031	0.2187	0.5429	1	0.4965	0.1201	0.3339	0.2702	0.3721	0.1483	0.0979	-0.1999	0.7464	0.5923	-0.3014	0.8274	0.5914	0.8444
TS	0.2102	0.2304	0.2223	0.2371	0.4965	1	0.0462	0.0019	0.0263	-0.0239	0.0376	0.1384	-0.0339	0.0321	0.8593	-0.0988	0.6079	0.2160	0.5449
ORB	-0.0449	-0.0794	-0.0383	-0.2827	0.1201	0.0462	1	0.7524	0.9023	-0.5001	-0.3463	0.6379	0.1179	-0.1707	0.0879	-0.4821	-0.0855	0.1151	-0.0281
DRB	0.1735	0.0518	-0.0693	-0.1321	0.3339	0.0019	0.7524	1	0.9621	-0.3836	-0.3277	0.6013	-0.0342	0.1134	-0.0051	-0.5718	0.0557	0.2977	0.1438
TRB	0.0974	0.0054	-0.0605	-0.2042	0.2702	0.0263	0.9023	0.9621	1	-0.4556	-0.3594	0.6552	0.0285	0.0066	0.0358	-0.5827	0.0056	0.2490	0.0884
AST	0.3101	0.0852	0.0538	0.4010	0.3721	-0.0239	-0.5001	-0.3836	-0.4556	1	0.4520	-0.4474	0.3503	0.4170	0.0972	0.1444	0.3583	0.2105	0.3500
STL	0.1717	-0.0073	-0.0109	0.1951	0.1483	0.0376	-0.3463	-0.3277	-0.3594	0.4520	1	-0.2968	0.2491	0.0366	0.0621	-0.0771	0.1208	0.2109	0.1681
BLK	-0.0405	-0.0346	-0.0969	-0.2349	0.0979	0.1384	0.6379	0.6013	0.6552	-0.4474	-0.2968	1	0.1172	-0.1871	0.0315	-0.4913	-0.1183	0.1464	-0.0434
TOV	-0.1256	0.0695	-0.1707	-0.2201	-0.1999	-0.0339	0.1179	-0.0342	0.0285	0.3503	0.2491	0.1172	1	-0.3182	-0.2220	-0.1150	-0.2122	-0.0650	-0.1879
USG	0.6196	0.0323	0.1380	0.4952	0.7464	0.0321	-0.1707	0.1134	0.0066	0.4170	0.0366	-0.1871	-0.3182	1	0.0784	-0.0687	0.5275	0.4047	0.5487
ORtg	0.2519	0.1677	0.2831	0.3447	0.5923	0.8593	0.0879	-0.0051	0.0358	0.0972	0.0621	0.0315	-0.2220	0.0784	1	-0.0581	0.7258	0.2694	0.6545
DRtg	-0.2657	-0.1174	-0.1090	-0.0222	-0.3014	-0.0988	-0.4821	-0.5718	-0.5827	0.1444	-0.0771	-0.4913	-0.1150	-0.0687	-0.0581	1	-0.1011	-0.7173	-0.3261
OWS	0.5982	0.0449	0.3878	0.6612	0.8274	0.6079	-0.0855	0.0557	0.0056	0.3583	0.1208	-0.1183	-0.2122	0.5275	0.7258	-0.1011	1	0.5243	0.9558
DWS	0.6262	0.0773	0.4890	0.6676	0.5914	0.2160	0.1151	0.2977	0.2490	0.2105	0.2109	0.1464	-0.0650	0.4047	0.2694	-0.7173	0.5243	1	0.7511
WS	0.6778	0.0606	0.4704	0.7427	0.8444	0.5449	-0.0281	0.1438	0.0884	0.3500	0.1681	-0.0434	-0.1879	0.5487	0.6545	-0.3261	0.9558	0.7511	1

Table 3.1. Correlation matrix

Statistical summary for selected variables

Variable	Salary	Games	Minutes	PER	ORB	USG	OWS	DWS	WS
Count	126	126	126	126	126	126	126	126	126
Min	0.1114	40	393	6.3	1	9.3	-0.9	0	-0.1
25%	2.3852	62	1290.25	12.425	2.125	16.425	0.6	0.9	1.725
50%	4.5	72.5	1859	14	3.7	19.1	1.65	1.6	3.15
75%	9.5914	79	2405	16.75	7.3	23.275	3.3	2.475	5.575
Max	22.9705	82	3122	29.8	17.5	33	14.8	6.4	19.2
Mean	6.7842	69.4603	1809.0159	14.9643	5.1548	19.9897	2.3373	1.8397	4.1825
Variance	31.8989	130.7144	494998.7997	18.5444	14.6339	24.6224	6.967	1.3893	11.5786
Standard Deviation	5.6479	11.433	703.5615	4.3063	3.8254	4.9621	2.6395	1.1787	3.4027
Skewness	1.1733	-0.7999	-0.1633	0.9653	1.0331	0.4215	2.105	0.9742	1.6949
Kurtosis	0.5069	-0.3825	-0.9483	1.2729	0.0725	-0.3821	5.4904	0.9371	3.5616

Table 3.2

3.1 Exploratory analysis for SALARY

As SALARY is the dependent variable that needs to be predicted, it is worth analysing it more carefully.

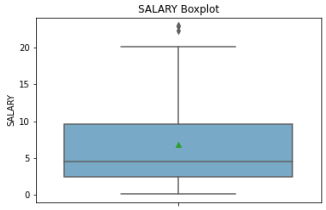

Figure 3.2

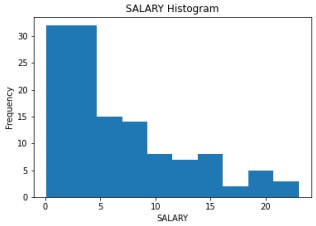

Figure 3.3

From the training dataset, there are 126 observations with significant difference in the lowest and highest salary of NBA players at 22.8591 million USD (i.e., 22.9705 – 0.1114). The mean and standard deviation are 6.7842 and 5.6479 million USD, respectively. Furthermore, 25, 50, and 75 percentiles are observed at 2.3852, 4.5, and 9.5914 million USD, respectively. This leads to interquartile range of 7.2062 million USD.

SALARY is not symmetrically distributed but considerably right-skewed, evident from the histogram and boxplot presented above. This is also supported by skewness of 1.1733. Overall, the dataset has some potential outliers at the end, indicating some players receive very high income comparing to the rest of the NBA.

3.2 Exploratory analysis for selected variables

There are many variables in the dataset and incorporating all of them into the report would be non-professional in presentation. Therefore, although all variables have been carefully analysed, a subset that has considerable relationships with SALARY is presented here to convey main ideas and enhance understandability and interpretability. These variables are Games, Minutes, PER, ORB, USG, OWS, DWS, and WS. They have some clear patterns of associations with SALARY, whereas for other predictors the data behaviours are not obvious and harder to interpret.

All variables presented here have some considerable degrees of relationship with SALARY at different levels. Some of them are evidently linear (e.g., Minutes, PER, WS), whereas the others indicate curved and nonlinear patterns in graphs (e.g., ORB). Also, not all variables from table 3.2 are symmetrically distributed, and there are both evidence for left-skewed (e.g., Minutes) and right-skewed (e.g., PER) shapes. Moreover, all variables are quite volatile with significant standard deviations such as Minutes (703.5615). In terms of kurtosis, no abnormally large absolute value is found, so variables might be acceptably bounded.

Regarding correlation, all factors signal a positive relationship with SALARY except ORB. This is an interesting finding from data since intuitively good players are usually expected to be well-rounded in different aspects, including offensive rebound. Furthermore, from the correlation matrix, these potential predictors have considerable associations with one another, so it should be cautious to select variables that do not overlap much information. Still, overall, all variables provide good sources for estimating and explaining SALARY. In other words, the data is quite rich and diverse, so the important task now is trying to learn as much as possible from it.

4. Variable selection, methodology, and modelling

4A. LINEAR REGRESSION

4A.1. Simple multiple linear regression model

It should be noted that models proposed would be compared based on k-fold cross validation technique, a common process to give an expectation of how well a model can perform on independent data (James et al. 2021, p. 198). Also, the report would adopt the popular approach of setting k=10 (James et al., 2021, p. 203). The notation would be cv_RMSE.

The report initially aims to find a multiple linear regression with already existing variables given (i.e., no transformation) to test the performance of such a simple model. Since the correlation suggests the linear relationships between two variables, Games, Minutes, PER, USG, OWS, DWS, and WS from the selected set above are all potential to be chosen. However, PER seems to be significantly correlated with OWS, DWS, and WS, so issue of multicollinearity may occur. Due to that reason, Games, Minutes, USG, OWS, DWS, and WS would be considered first.

Nevertheless, there are reasons why OWS, DWS, and WS should not be fitted together. Logically, from the glossary, these 3 variables seem to represent the same factor: the contribution to wins of a player measured as percentage (NBA, 2021). Intuitively, a player plays well on one day should have good impacts on the team either offensively and defensively, which in turn would correspond to high OWS, DWS, and WS. Statistically, this idea is further supported by the fact that correlations between them are considerably high (see table 3.1). Therefore, including all 3 variables cause multicollinearity:

Proposed population relationship	$\boldsymbol{SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG + \beta_4 OWS + \beta_5 DWS + \beta_6 WS + \varepsilon}$					
VIF	2.2559	3.915	1.6408	2600.8901	514.8588	4327.0619

Table 4A.1

As a result, only one of OWS, DWS, and WS would be selected, and it would be compared by cross validation RMSE:

Proposed population relationship	**cv_RMSE**
$SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG + \beta_4 OWS + \varepsilon$	3.7867
$SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG + \beta_4 DWS + \varepsilon$	3.5651
$SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG + \beta_4 WS + \varepsilon$	3.6586

Table 4A.2

As an outcome, the second model in table 4A.2 would be chosen. Finally, adding PER to that model increases cv_RMSE to 3.6, so it is decided that $SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG + \beta_4 DWS + \varepsilon$ would be kept. Some main statistics are presented below:

$$\boldsymbol{\widehat{SALARY} = 2.1032 - 0.1721 Games + 0.0039 Minutes + 0.3212 USG + 1.7149 DWS}$$

Adjusted R-squared	0.617
R-squared	0.629
AIC	677.8
BIC	692
Training RMSE	3.4947
All coefficients' estimates except β_0 are significant at α=5% (see appendix a)	

Table 4A.3

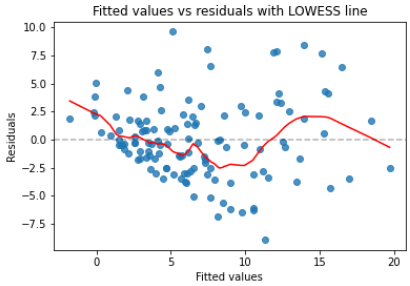

Figure 4A.1

4A.1.1. Strength of fit

$R^2_{adjusted}$= 0.617 shows a moderate strength of fit to the data. Moreover, training RMSE = 3.4947 million USD is the standard deviation of residuals, but it may be necessary to consult with experts to determine whether such an error is acceptable when predicting salary. However, overall, the model predicts the NBA's salary quite acceptably.

4A.1.2. Goodness of fit

The LOWESS line of residuals against the fitted values suggests quite poor goodness of fit. Particularly, there might be nonlinearity, with over-predicting salary occurs at 2 million USD < $\widehat{SALARY}$< 12 million

USD and under-predicting salary occurs at 12 million USD < $\widehat{SALARY}$ < 18 million USD. Overall, the line fluctuates, so the assumptions in linear regression models about linearity (i.e., $E(Y_{i,j}|X=j)=\mu_j$) and exogeneity (i.e., $E(\varepsilon_{i,j}|X=j)=0$) may not be satisfied. Regarding the existence of 4th finite moment, all covariates by construction should be bounded, so the assumption might be satisfied here. The variance seems to widen as fitted values get larger, so homoskedasticity seems to be violated. Next, whether the data is independent may not be obvious here, but dependence could occur since the allocation of minutes played, usage, or win shares is likely to be related between players in the same club. However, without further evidence or description, no conclusion is made, and the independence is broadly accepted here. Finally, from the correlation matrix, the assumption about no perfect collinearity is acceptable.

In short, although the model can moderately explain the variations in SALARY, some of the underlying assumptions may not be well valid. As a result, further analysis is conducted to improve usefulness of the model.

4A.2. Variable transformation

In addition to cv_RMSE, here R-squared adjusted, AIC, and BIC would also be considered to give a better idea for client. In this section, some techniques in features engineering are utilised to transform existing or unadded variables to improve the model, which include:

1. Taking exponential of USG

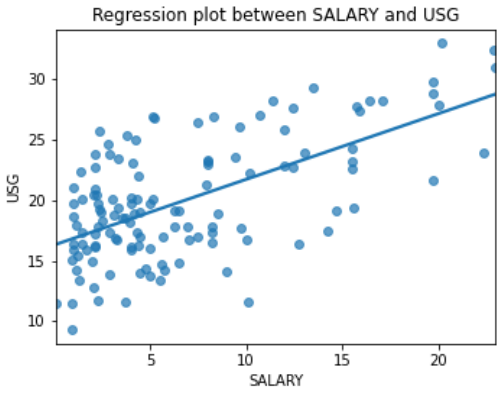

Figure 4A.2

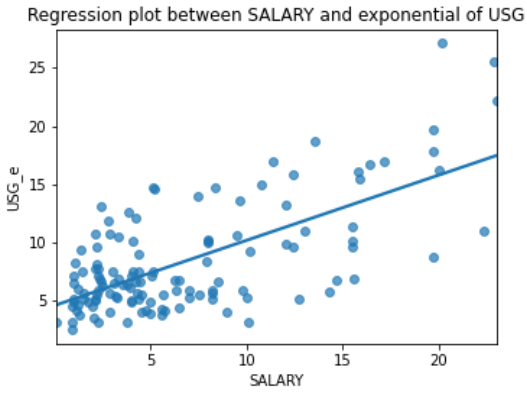

Figure 4A.3

Proposed population model: $\boldsymbol{SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG\_e + \beta_4 DWS + \varepsilon}$

cv_RMSE	R-squared adjusted	AIC	BIC
3.4658 (↓)	0.641 (↑)	669.8 (↓)	684 (↓)

Table 4A.4

Despite good correlation between USG and SALARY already, the pairplot suggests potential pattern of an exponential function, so attempt to transform USG has been made. It should be noted that since USG is on a quite large scale, the variable has been reduced 10 times prior to transformation for better visulisation. Replacing USG_e (i.e., exponential of USG), the new model becomes better regarding cv_RMSE, R-squared adjusted, AIC, and BIC. Therefore, multiple linear regression is updated accordingly from now on.

2. Log transformation for ORB

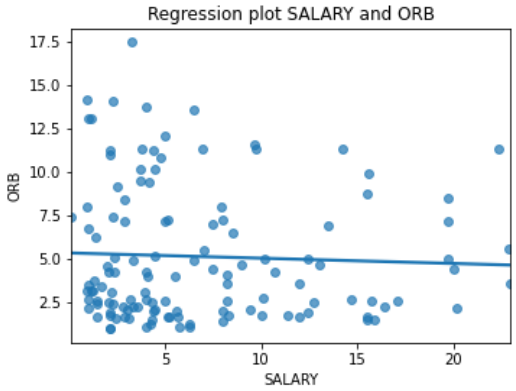

Figure 4A.4

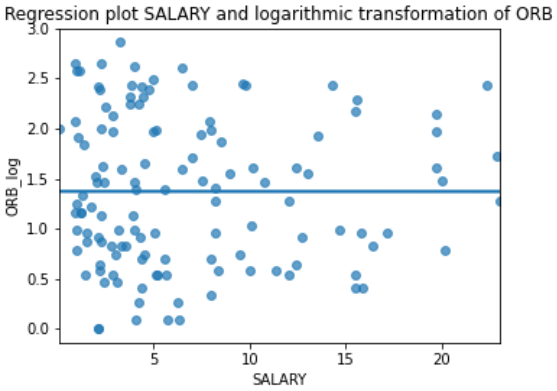

Figure 4A.5

Proposed population model: $\boldsymbol{SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG\_e + \beta_4 DWS + \beta_5 ORB\_log + \varepsilon}$

cv_RMSE	R-squared adjusted	AIC	BIC
3.4155 (↓)	0.656 (↑)	665.2(↓)	682.2 (↓)

Table 4A.5

The pairplot suggests potential pattern of a logarithmic function, so attempt to transform USG has been made. Adding ORB_log (i.e., log of ORB), the new model becomes better regarding cv_RMSE, R-squared adjusted, AIC, and BIC comparing to the updated one just above. Therefore, multiple linear regression is updated accordingly from now on.

$SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG\_e + \beta_4 DWS + \beta_5 ORB\_log + \varepsilon$ now would be the chosen as the final proposed population model. Some main statistics are presented below:

$$\widehat{\boldsymbol{SALARY}} = \boldsymbol{3.7419 - 0.1845 Games + 0.0048 Minutes + 0.4062 USG\_e + 1.1304 DWS + 1.2305 ORB\_log}$$

Adjusted R-squared	0.656
R-squared	0.67
AIC	665.2
BIC	682.2
Training RMSE	3.3114
All coefficients' estimates except $\boldsymbol{\beta_0}$ are significant at α=5% (see appendix b)	

Table 4A.6

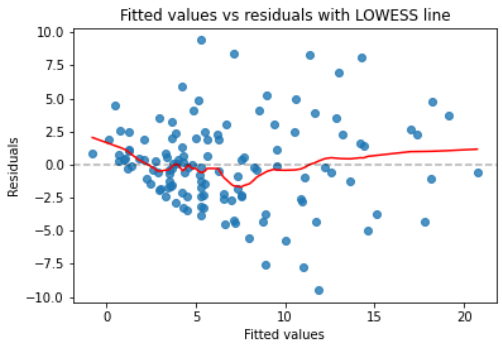

Figure 4A.6

4A.2.1. Strength of fit

$R^2_{adjusted}$= 0.656 shows a stronger performance comparing to that of the pre-transformed model. Also, SER = 3.3114 million USD has decreased as well, but it may still be necessary to consult with experts to determine whether such an error is acceptable when predicting salary. However, overall, the model predicts the NBA's salary quite well.

4A.2.2. Goodness of fit

Despite existing under- and over-prediction in the same range as discussed above, the LOWESS line demonstrates significantly increase in goodness of fit, suggesting that the transformations may have represented the relationships better. The line is relatively flat, so linearity and exogeneity assumptions might be satisfied. Furthermore, similarly to before, 4th finite moment of variables can be accepted, and no perfect collinearity assumption is still acceptable. Nevertheless, homoskedasticity and independent of data continues to be questionable. Still, overall, the goodness of fit has generated more confidence in the outputs produced by the model.

4A.2.3. RMSE test set

The performance on the test set is reflected via RMSE test set of 4.057 million USD, and this is below the criteria put forth by the client.

4A.2.4. Final consideration

Multicollinearity: One common issue when fitting multiple linear regression model is potential collinearity between predictors may occur. Fortunately, variance inflation factors calculated show no particular figure larger than the common benchmark of 5 (James et al., 2021, p.102). Furthermore, the average is well below 5 at 2.4492, so multicollinearity is not a concerned issue here, thus generating more confidence in the estimates.

Proposed population relationship	$\boldsymbol{SALARY = \beta_0 + \beta_1 Games + \beta_2 Minutes + \beta_3 USG\_e + \beta_4 DWS + \beta_5 ORB\_log + \varepsilon}$				
VIF	2.3376	4.5204	1.5836	2.3469	1.4576
Average	2.4492				

Table 4A.7

Omitted Variable Bias: This issue requires that the omitted variable must (1) impact the response and (2) be related to other the explanatory variables (Fox, 2016). However, both cannot be reliably observed and ascertained in the dataset. Furthermore, it is very sensitive to conclude whether there is a determinant relationship instead of just correlation. Further analysis about the concern deviates from the main objective of the report, hence it would not be conducted. Still, the idea is presented to remind client to have reasonable consideration when interpreting the outcome.

4B. KNN REGRESSION

4B.1. Discussion

There are several problems in KNN model that needs considering. Firstly, KNN regression suffers from curse of dimensionality. Specifically, this means that as the number of predictors increases, there would be too many dimensions that it is unlikely for an observation to have a real nearby neighbour (James et

al., 2021, p. 108). The effect would become more severe if there are only a few observations due to cost and other constraints (Li, 2009). Indeed, with limited 126 observations in the training data, this is a significant concern here. Therefore, it is cautious regarding how many predictors to be included in the model. A common rule is that number of predictors should be equal or less than the square root of total observations (Li, 2009), so it is expected that there should be less than $\sqrt{126} = 11.22$ predictors. With the issue of dimensionality, however, the number may be even lower. Secondly, another problem is the computational cost of the model. Particularly, KNN regression requires storing of training data over time, and the calculations rocket as more predictors are introduced (James et al., 2021). This is because several iterations must be conducted to find the optimal neighbours. Finally, the model is nonparametric, so it requires no assumptions regarding behaviours of data. While this fact makes KNN more robust, there is no clear starting point to select appropriate predictors.

All arguments above make it challenging to choose a good KNN regression model. Without those constraints, the task would have been done by just applying best subset selection. Hence, the report proposes a systematic approach to overcome the difficulties. The aim would be to try providing distinctive information as much as possible to the model with limited number of predictors. In other words, predictors that seem to give the same information should not be added together since it increases the complexity without contributing significantly new knowledge to the model. As a result, in this approach, predictors would be grouped according to their similarity, and one representative would be selected. As aforementioned, cv_RMSE would be the basis for comparison. To sum up, the method wants to avoid three problems above with a trade-off of not having the optimal model comparing to best subset selection. Still, the final choice would need to satisfy the performance criteria set forth by the client (i.e., RMSE test set < 4.1 million USD).

It should be noted that for each set of variables proposed, a designed function has been developed to find K, the optimal number of neighbours for that set through multiple iterations. In the supporting workings attached, however, it should be noted only discrete values of K from 1 to 50 (not 1 to 125) would be considered to avoid curse of dimensionality issue and unnecessary complexity.

4B.2. Variable selection

1. WS – DWS – OWS – PER – USG – TS – ORtg - DRtg : According to the glossary, these 8 predictors seem to indicate the overall performance and impact of players to their team (NBA, 2021). For example, PER demonstrates how efficient a player is, and USG describes how a player effectively utilises team plays (NBA, 2021). Evidently, correlations between these variables are also considerable though the relationships may be complicated and incompletely clear (see table 3.1 and appendix c). Of these variables, WS, PER, DWS indicate the evident strong linear relationships with SALARY, whereas the others may present some non-linear or unclear patterns. As the report is data-driven, cv_RMSE would be compared:

Predictors	cv_RMSE	K
WS	4.29	22
DWS	4.4509	18
OWS	4.5979	17
PER	4.3361	15
USG	4.3107	12
TS	5.5541	42

ORtg	5.4493	29
DRtg	5.4766	50

Table 4B.1

Therefore, WS would be chosen, and the model is updated accordingly

2. Minutes: This variable is separately grouped because it does not belong to a particular attribute of player's ability. Evidently, correlation and scatter plot between Minutes and SALARY suggests a strong linear relationship, so it is expected that this can be a good predictor. Intuitively, it makes sense since highly paid players should be superstars of teams and be allocated large proportions of gameplay. As the report is data-driven, cv_RMSE would be compared:

Predictors	**cv_RMSE**	**K**
WS, Minutes	4.2531(↓)	24

Table 4B.2

The cv_RMSE has decreased comparing to the previous updated model, so this suggests Minutes as good explanation of SALARY. Therefore, the model is updated accordingly.

3. TOV – AST: According to the glossary, these 2 predictors seem to indicate how good players are in offence specifically (NBA, 2021). For example, TOV demonstrates player's ability to avoid losing ball in an offensive play, and AST (assist) describes how a player creates opportunities for other players to score (NBA, 2021). Evidently, correlations between these variables are also considerable though the relationships may be complicated and incompletely clear (see table 3.1 and appendix c). Of these variables, there are not clear patterns of associations with SALARY, so they may not be good predictors in this study design. However, As the report is data-driven, cv_RMSE would be compared:

Predictors	**cv_RMSE**	**K**
WS, Minutes, AST	4.4579(↑)	17
WS, Minutes, TOV	4.3507(↑)	16

Table 4B.3

Therefore, both would not be added, and the set of predictors accepted are still [WS, Minutes]

4. ORB – TRB - DRB – BLK – STL: According to the glossary, these 6 predictors seem to indicate how good players are in defence (NBA, 2021). For example, ORB, TRB, and DRB demonstrates player's ability to rebound (i.e., take the ball back for their team) in case of unsuccessful shooting, and BLK describes how a player prevents opponents from scoring (NBA, 2021). Evidently, correlations between these variables are also considerable though the relationships may be complicated and incompletely clear (see table 3.1 and appendix c). Of these variables, ORB demonstrates probable logarithmic relationship with SALARY, whereas the others may present some unclear patterns that cannot be identified easily. As the report is data-driven, cv_RMSE would be compared:

Predictors	**cv_RMSE**	**K**
WS, Minutes, ORB	4.1672(↓)	8
WS, Minutes, TRB	4.2955	9
WS, Minutes, DRB	4.3371	7
WS, Minutes, BLK	4.3602	12
WS, Minutes, STL	4.3756	20

Table 4B.4

Therefore, ORB is selected, and the model is updated accordingly.

5. Age - Games: This variable is separately grouped because it does not belong to a particular attribute of player's ability. Furthermore, these two predictors are separated from Minutes because they have poorer correlations with SALARY comparing to Minutes, suggesting non-linear relationships. Indeed, they are two of some predictors that have the lowest correlations with the response. In other words, Minutes is considered first due to its evident relationship with SALARY already. As the report is data-driven, cv_RMSE would be compared:

Predictors	cv_RMSE	K
WS, Minutes, ORB, Age	3.924	7
WS, Minutes, ORB, Games	3.8219(↓)	7

Table 4B.5

Therefore, Games is selected, and the model is updated accordingly.

4B.3. RMSE test set

The performance on the test set is reflected via RMSE test set of 3.7705 million USD, and this is significantly below the criteria put forth by the client.

4B.4. Final consideration

- KNN regression is a nonparametric model, so there are no assumptions to be checked
- With only 126 observations in the training data, the optimal k = 7 is not significantly large, and this aligns with several studies in the field when data are limited (Li, 2019). When more data are gathered, a bigger k may be expected.

4C. KERNEL REGRESSION

4C.1. Discussion

The report presents kernel regression with Nadaraya-Watson estimator as the third model to estimate SALARY. This is a nonparametric method that offers some advantages over KNN regression. Specifically, in KNN, weights assigned are equal among all y_i that have x_i close to x (James et al., 2021, p. 105). However, for kernel regression with Nadaraya-Watson estimator, a slight change is made when y_i that has x_i more nearby to x would be given a higher weight (Portugues, 2021). The aim is to have a better prediction that is as close to the true value as possible. Still, as this method is nonparametric, one drawback is that there is no obvious starting point for variable selection. Therefore, the report would base on predictors in multiple linear regression and KNN regression to have some initial choices. Indeed, because the motivation under kernel regression is very similar to that of KNN, it is expected that both should have somewhat identical variables.

It should be noted that on this section, the training data set is further split to 2 parts, one for training, and one for validation. The ratio would be the common 80/20 used in the literature (Gholamy, Kreinovich, and Kosheleva, 2021), and random state of 1 is set according to client's requirement. Therefore, the validation set RMSE would now be the benchmark for comparison.

4C.2. Variable selection

As observed above from the pairplot, the correlation of Minutes and SALARY is quite significant, so a relationship can be relatively ascertained. Furthermore, Minutes is chosen as a reliable predictor for both linear and KNN regressions above, thus suggesting its validity. Evidently, Minutes is selected as the first independent variable in the model.

Games is another considerable predictor. This variable's correlation with SALARY is weaker than Minutes, but it is still contributing a significant level of variation explanation in either linear or KNN models. As a result, Games is a reliable candidate for predictor in kernel regression.

Since the report is data-driven, validation set RMSEs would now be compared:

Predictors	**Validation set RMSE**
Games	4.7762
Minutes	4.1477
Games, Minutes	3.7983

Table 4C.1

Therefore, introducing either Games or Minutes later in the model does decrease the validation set RMSE, so two predictors would be chosen, and the model is updated accordingly.

In two previous models, DWS is chosen in linear regression, whereas WS is selected for KNN. Therefore, although they are relatively similar in terms of meanings interpretation, there is a disagreement in variable selection. However, it indicates that predictors relating to win shares play an important role in estimating SALARY of NBA players. With this regard, OWS, DWS, and WS are analysed next. Validation set RMSEs are now compared:

Predictors	**Validation set RMSE**
Games, Minutes, WS	4.0002
Games, Minutes, DWS	3.1879 (↓)
Games, Minutes, OWS	4.3865

Table 4C.2

Therefore, DWS would be chosen, and the model is updated accordingly.

As argued that KNN and Nadayara-Watson kernel regression have the same underlying motivation and lack clear starting point to select variables, the report now would follow similar approach conducted for KNN regression. More specifically, only predictors that potentially bring new information should be considered. In this case, as DWS has incorporated information about overall performance and defensive ability of NBA players to some extents (NBA, 2021), only Age and variables relating to offence (i.e., TOV and AST) are conducted. In other words, evident from validation set RMSE of only 3.1879, kernel regression has performed exceptionally well, so there might be not necessary to analyse all variables that carry partially similar information with one another. In this case, focusing on those that are considerably different from the existing predictors Games, Minutes, and DWS can be a better practice in terms of computational and analysis costs.

Validation set RMSE would still be the main basis for comparison:

Predictors	Validation set RMSE
Games, Minutes, DWS, AST	3.3422
Games, Minutes, DWS, TOV	3.1809 (↓)
Games, Minutes, DWS, Age	4.3778

Table 4C.3

Therefore, TOV would be chosen, and the model is updated accordingly.

4C.3. RMSE test set

The performance on the test set is reflected via RMSE test set of 4.0215 million USD, and this is below the criteria put forth by the client.

4C.4. Final consideration

- Kernel regression is a nonparametric model, so there are no assumptions to be checked

5. Conclusions and analysis

Since the NBA league's requirement indicates the main aim of predicting SALARY, based on RMSE test set, KNN regression could be the best choice. However, there are at least four findings that should be acknowledged:

1. The choice of KNN regression does not completely undermine the power of Nadayara-Watson kernel regression, and there are at least three reasons. Firstly, the chosen kernel model with four non-transformed variables has already outperformed multiple linear regression and satisfied the RMSE test set criteria. Secondly, the approach of selecting variables here might not be entirely effective for kernel method. In other words, there may be another better set of predictors, but it requires another independent, intensive, and careful report to find out. Thirdly, the training set is already small, so splitting it up further might not give enough valuable information for the kernel model to study.
2. Multiple linear regression should not completely be undermined. To begin with, it should be acknowledged that model's restrictive characteristic can be one of reasons causing poor performance. More specifically, under linear regression, all variables are selected with the frame that the relationship must be linear along with assumptions such as homoskedasticity or exogeneity. However, the reality is that relationships might be in more complicated forms, and some assumptions cannot be examined with certainty. In contrast, KNN and Nadayara-Watson kernel regression do not constrain their views from linearity only, and that may partly explain why they perform predictions considerably better. Still, these two models do not explicitly explain the relationships between predictors and the response. On the other hand, linear regression requires estimates of parameters which are good benchmarks to analyse how the independent and dependent variables are associated. To sum up, there is a trade-off between predictability and interpretability here.
3. Through analysis, the report has found several attributes of players that are important explanations for NBA players' salary as an answer for initial questions put forth at the 'business context' section. It turns out that Games, Minutes, and variables relating to win shares are good indicators to estimate an appropriate salary. While there is non-alignment in the final set of predictors chosen, these mentioned variables are always considered, thus suggesting their validity. Also, based on the prediction outcome and coefficient estimates, the hypothesis that players assigned much time with significant contribution tend to have higher salary may be correct.

4. As mentioned, the choice of KNN signals the requirement of storing training data for future prediction, so storage cost is a limitation that needs considering here. Also, only 126 observations may be said to be very limited. Therefore, some assumptions under linear regression cannot be addressed with certainty, and an observation may not have lots of "true" nearby neighbours to utilise KNN and Kernel regressions fully. Whenever possible, it is recommended that NBA league collect more data.

PART B

1a

To begin with, when fitting a model, researchers try to find a good explanation for behaviours and patterns of data that was collected. If sophisticated techniques are utilised, it is possible that eventually the variations in observed data can be well-explained. However, that outcome is only desired if the objective is just to have a good description about the collected sample. Unfortunately, the main aim in most studies is to predict outcomes in different situations based on the data gathered (James et al., 2021). More specifically, uncertainty of the future unknowns is the main consideration, not the already known data. With this regard, what researchers really focus on is to select the appropriate model explaining variations in the sample that is as close to the true population relationship as possible. Trying to capture everything would cause overfitting which means that the model seeks to explain natural random variations in sample as well, and this does not reflect behaviours of the population (James et al., 2021).

To overcome overfitting, because the true population model is generally not known or mis-specified in training data (Maldonado & Greenland, 1993), the best approach might be to see how well the proposed model predicts compared to true values on a different dataset. Still, this testing data is not readily available all the time (James et al., 2021). This may be because the difficulties in collecting procedure or cost concerns (Li, 2019). As a result, the available set is usually partitioned into training set and validation set. The training data would then be used as main inputs for conducting explanatory analysis and selecting relevant models for fitting. Furthermore, this training set would also be based on to estimate coefficients. On the other hand, the validation set acts like testing data, which is to understand how model performs on new unknown information. If the errors on the validation set prediction is within acceptable limit, then it suggests some validity of the proposed relationship. Otherwise, when poor prediction on validation set occurs, researchers may have faced the issue explained above: overfitting. As a result, rethinking about appropriate models and predictors may be required. In short, the validation set is helpful in the model selection process to help researchers point out whether the approaches used may be correct or not (Hastie, Tibshirani, and Friedman, 2009, p. 222).

On final note, validation set does not replace test set. Over time when the data is rich and testing sample becomes available, it should be utilised to check the generalisation of error on the chosen model (Hastie, Tibshirani, and Friedman, 2009, p. 222).

1b

The random partition of data has been recommended in the field (Joseph and Vakayil, 2019), and the main reason is non-random data causes biases in estimates. To begin with, if researchers intentionally choose observations with specific characteristics for the training, the data used to fit the model would be subjective towards those selected, whereas the left-out data are not considered. In other words, too much information of specific groups of similar attributes in the population are given, but there is not

sufficient knowledge from the others for the model to study. Consequently, overfitting is likely to occur, and the model may be tied to explain variations in selected observations only rather than the overall population pattern. Furthermore, intentional selection leads to biased estimates of errors to be either larger or smaller as well. Therefore, this would give false perception of the true performance of the model and leads to poor choice of method as well as interpretation of predicted figures. Indeed, attempting to subjectively choose data for training without logical reasons has been considered as an unethical practice (Hand, 2018). Unfortunately, this is not rare in the field (Martin and Mike, 2013).

In a less extreme situation, standard split, which is to divide the data based on the order already given, may still present unbiased issues (Gorman and Bedrick, 2019). Particularly, the intuition may be that as data has already been independently and randomly gathered during collection process, no further randomisation is needed. Nevertheless, there are two potential flaws. Firstly, even the process has been designed to collect arbitrary data, biases can still be found in the way it is implemented. For example, collectors may be biased towards specific location, gender, or timeframe (Royal Geographical Society, n.d.). Consequently, the data may be unconsciously presented in an implicit systematic way such as figures of the same region are nearby one another. Secondly and more importantly, Gorman and Bedrick (2019, p.1) have found evidence that outcomes from standard split still have patterns that cannot be reproduced with randomly generated data, thus suggesting biases. In short, the gathered data through random collection process may not be truly arbitrary, so subjective problems might persist.

To sum up, random partition ensures that there is no bias in prediction or interpretation, so the training set and validation set can be good inferences of the population distribution and behaviour. Indeed, researchers want to gain insight on the population patterns and not predictable outcomes that are due to subjective selection of data.

2a

If simple linear regression is used, then the log-scale should be preferred. There are several reasons for this approach. To begin with, the main motivation and assumption of this model is the relationship between dependent variable Y and independent variable X can be linearly approximated (James et al., 2021, p.61). Therefore, although the true model may not be known (Maldonado & Greenland, 1993, p.1), it is reasonable that the pattern of the data must somehow indicate a straight-line relationship. In this case, log-scale shows a behaviour that is closer to linearity comparing to the normal one. As result, log data fits better with the model's goal. In short, a model should only be considered if there are evidence of its existence; otherwise, there are no logical and meaningful justification to do so.

Furthermore, if analysts would insist on using normal scale for simple linear regression, homoskedasticity assumption of the model may also be violated. More specifically, the nature of a straight line is unable to capture the curved pattern between Fertility and PPgdp, so there might be ranges of the independent variable where the residuals become either larger or smaller. However, more importantly, the residuals change dramatically due to underfit would cause the standard errors to be large. Consequently, under the model construction, the estimates of the parameters are not reliable. The confidence intervals are wider and may even include 0, so predicted outputs must be interpreted with caution.

In contrast, since the log-scale data behaves more closely to a linear line, it can overcome issues analysed above. Particularly, although the constant variance assumption cannot be ascertained, they

may be generally accepted. Also, standard errors are expected to be relatively small, thus generating more confidence in the coefficient estimates. Therefore, log-scale may be a more reasonable choice to fit here.

On final note, researchers must be more careful in interpreting the relationship between two variables, and sometimes back transformation to the normal variables may be needed for enhanced understanding.

2b

Given the situation and limited information given, there are reasons to remove Purban from the model, one of which is based on its p-value. More specifically, at typical cutoff 5% (James et al., 2021, p.68), p-value of 0.063 indicates that it may be unlikely due to chance to observe this association between Purban and logFertility in the presence of any real relationship. Therefore, it is not statistically significant to conclude an association between two variables, so keeping Purban in the model may not yield trustworthy prediction. In other words, there is evidence that these variables do not have relationship, so there might be no point trying to explain variations in one based on the other. Furthermore, under statistical perspective, Purban may be said to negatively affect other parameters as well. Particularly, with only 193 observations, the dataset can be argued to be relatively small comparing to common sets observed in the field (ScienceDirect, n.d.), so adding non-significant predictor may unnecessarily cause other estimates to experience wider confidence intervals due to less degrees of freedom. In short, introducing a predictor that may not have significant explanation power as well as affect certainty in other statistics may not be a reasonable practice.

Nevertheless, it should not be ascertained to say Purban is completely not useful. Firstly, although 5% level of significance is a usual choice, it is not a requirement to conform. Based on the goal of study, if researchers can reasonably justify a different cutoff such as 10%, Purban may still be accepted. Secondly, if typical 5% is kept as the benchmark, the p-value of 0.063 may not be highly different from the limit. This might indicate that the random nature of sample might have resulted in this outcome, and the real association between Purban and logFertility can still be considerable. Indeed, Altman and Krzywinski (2017) found that several factors regarding sample even with small magnitude can lead to a large difference in the p-value. The association between Purban and logFertility is further supported by studies in social science claiming that urbanisation leads to reduction in birth rates (Yi and Vaupel, 1989). Moreover, from multicollinearity perspective, observed variance inflation factor of $\frac{1}{1-0.78^2} = 2.5536$ is not problematic (James et al., 2021, p.102). Therefore, concluding an insignificant relationship quickly here can be undesirable. Thirdly, Purban coefficient may be unfavourable when the model is used for prediction, but for inference it might not be very problematic. Indeed, including the predictor in the model gives researchers a better understanding about its relationship with the response to gain more insight in the population behaviour. As a result, if predictability is not the main concern, it might not harm to include and study Purban.

In short, there is statistical evidence against the inclusion of Purban, but this cannot invalidate the predictor absolutely. Judgments based on data-driven findings are good, but they must be placed in the study context for reasonable decision-making. Therefore, it is recommended that more information be considered before removing. Also, if possible, collecting more data for studying to generate more confidence in the output is a good consideration.

References

Altman, N & Krzywinski, M 2017, 'Points of significance: P values and the search for significance' Nature Methods, vol. 14, no. 1, p. 3, doi: 10.1038/nmeth.4120.

Fox, J 2016, Applied regression analysis and generalized linear models, Third edition., SAGE, Thousand Oaks, California.

Gholamy, A, Kreinovich, V, and Kosheleva, O, 2018, *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*, viewed 21 September 2021, http://www.cs.utep.edu/vladik/2018/tr18-09.pdf

Gorman, K, and Bedrick, S 2019, 'We need to talk about standard split', Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2786–2791, https://aclanthology.org/P19-1267.pdf

Hand, DJ 2018, 'Aspects of Data Ethics in a Changing World: Where Are We Now?' Big Data, vol. 6, no. 3, pp. 176–190, doi: 10.1089/big.2018.0083.

Hastie, T, Tibshirani, R, & Friedman, J 2009, The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition , 2nd ed. 2009., Springer New York, New York, NY, doi: 10.1007/978-0-387-84858-7.

James, G, Witten, D, Hastie, T, and Tibshirani, R 2021, An Introduction to Statistical Learning with Applications in R, Second edition., Springer

Joseph, VR and Vakayil, 2021, 'SPlit: An Optimal Method for Data Splitting' Technometrics, AHEAD-OF-PRINT, pp. 1–11, doi: 10.1080/00401706.2021.1921037.

Li, S 2009, Random KNN modeling and variable selection for high dimensional data, ProQuest Dissertations Publishing.

Maldonado, G and Greenland, S 1993, 'Interpreting Model Coefficients When the True Model Form Is Unknown', *Epidemiology (Cambridge, Mass.),* vol. 4, no. 4, pp. 310–318, doi: 10.1097/00001648-199307000-00006.

Martin, C & Blatt, M 2013, 'Manipulation and misconduct in the handling of image data'*, The Plant Cell*, vol. 25, no. 9, pp. 3147–3148, doi: 10.1105/tpc.113.250980.

NBA 2021, *NBA Advanced Stats,* viewed 10 September 2021, https://www.nba.com/stats/help/glossary/https://www.nba.com/stats/help/glossary/

Portugués, G 2021, *Notes for Predictive Modeling*, viewed 21 September 2021, https://bookdown.org/egarpor/PM-UC3M/

Royal Geographical Society n.d., *A guide to avoid biased data*, viewed 18 September 2021, https://www.rgs.org/CMSPages/GetFile.aspx?nodeguid=becd49a7-a675-480a-bc70-64929df32f13&lang=en-GB

ScienceDirect n.d., *Large data set,* viewed 23 September 2021, https://www.sciencedirect.com/topics/computer-science/large-data-set

Yi, Z & Vaupel, JW 1989, 'The Impact of Urbanization and Delayed Childbearing on Population Growth and Aging in China' Population and Development Review, vol. 15, no. 3, pp. 425–445, doi: 10.2307/1972441.

Appendix

a) Additional information for table 4A.3

	coef	Std err	t	P>|t|
Intercept	2.1032	2.707	0.777	0.439
Games	-0.721	0.041	-4.228	0
Minutes	0.0039	0.001	4.765	0
USG	0.3212	0.078	4.135	0
DWS	1.17149	0.360	4.765	0

b) Additional information for table 4A.6

	coef	Std err	t	P>|t|
Intercept	3.7419	2.182	1.715	0.089
Games	-0.1845	0.040	-4.658	0
Minutes	0.0048	0.001	5.360	0
USG_e	0.4062	0.079	5.122	0
DWS	1.1304	0.385	2.937	0.004
ORB_log	1.2305	0.482	2.554	0.012

c) Regression plots for every pair of variable