QBUS2820

Assignment 1

TASK A: Money Ball
TASK B: Essay

500354333

Content

QBUS2820 Assignment 1

Task A

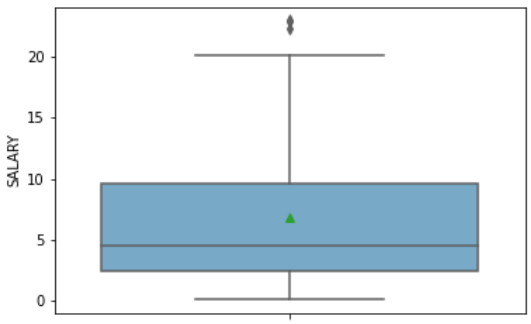
Introduction:

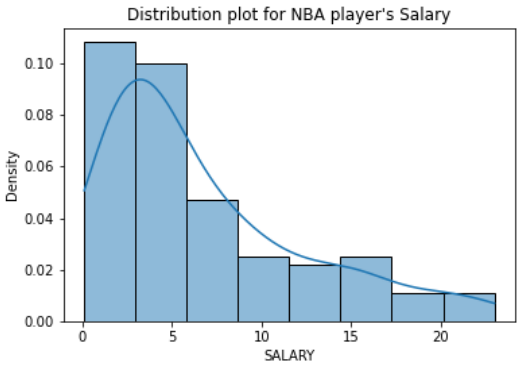
NBA is an abbreviation of National Basketball Association is the most popular basketball tournament in the world. According to the news, Stephen Curry who played for Golden State Warriors receive a whooping $45.780.966 millions followed by John Wall who played for Houston Rockets with $44.310.840 millions, the average salary for NBA 2020-2021 season was $7.5 millions. The given information show there is a huge salary gap between a "superstar" who being paid almost 700% more compared to the average player in the tournament, The purpose of this analysis is to identify the features that contribute or correlated with a NBA player salary, so we could predict a reasonable salary when signing a NBA player in the future based on their performance metrics. In the data set we are provided NBA players performance metric such as player position, player team, player age, games and minutes played, personal efficiency rating (PER), true shooting percentage (TS), offensive rebounds (ORB), defensive rebounds (DRB), total rebounds (TRB), assists (AST), steals (STL), blocks (BLK), the turnover percentage per possession (TOV), usage percentage (USG), offensive rating (ORtg), defensive rating (DRtg), Offensive win shares (OWS), defensive win shares (DWS) and win shares (WS). This model will also help to detect any underpayment or overpayment that NBA player received based on their on-court performance. In the future, we hope this model allows a fair and equal salary among all NBA players based on their performance.

Initial Data Analysis and Data cleaning:

We start from dropping any N/As that the data contains. From the initial data analysis (IDA) that we conduct, our training data contains 126 observations with 21 variables. From 21 variables, 2 of them are categorical variable ("Position" and "Teams") and the rest of them (19 variables) are numerical variables.

As we need to investigate NBA player salary as our response variable, we did numerical summary, plot histogram and boxplot on "SALARY" variable:





Statistical summary of train "SALARY"	
Count	126.000000
Mean	6.784165

Standard Deviation	5.647912
Minimum	0.111444
25%	2.385205
50%	4.500000
75%	9.591416
Maximum	22.970500

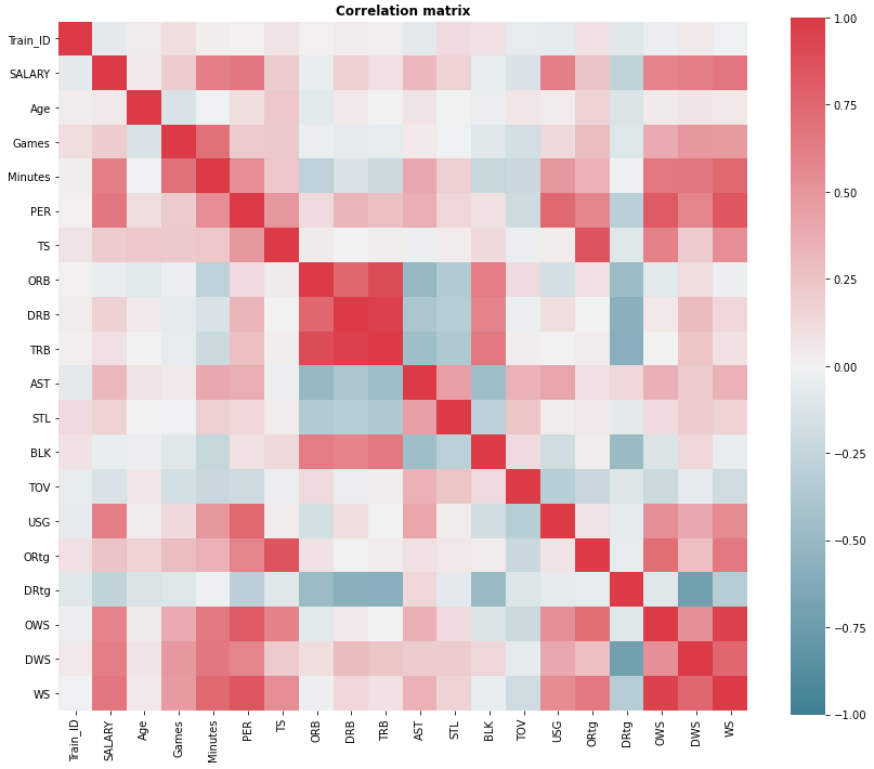
From the result shown above, the maximum salary of NBA player is 22.970500 millions and the minimum salary of NBA player is 0.111444 millions, this shown a significant gap of 22.859056. Additionally, the mean salary of NBA player is observed at 5.647912 millions and standard deviation of 5.647912 millions. Moreover, the 25, 50, and 75 percentiles being reported as 2.385285, 4.500000, 9.591416 respectively.

The histogram and boxplot above indicate the distribution of salary are right-skewed, and there is a data saturation in the range of 0 to 5 million, this indicate that most of NBA player earn between 0-5 million salary.

Following the IDA of "SALARY" variables, we produce a pairwise correlation table, this table show how strong "SALARY" and other 19 numerical variables correlate. The result in this table ranges from 0 to 1 for a positive correlation and 0 to -1 for a negative correlation. The stronger the correlation, the result will be closer to 1 for positive correlation and -1 for negative correlation.

Pairwise correlation matrix of SALARY vs numerical variables	
	SALARY
TRAIN_ID	-0.077319
SALARY	1.000000
Age	0.053676
Games	0.205137
Minutes	0.621254
PER	0.670230
TS	0.210184
ORB	-0.044877
DRB	0.173453
TRB	0.097399
AST	0.310137
STL	0.171666
BLK	-0.040450
TOV	-0.125629
USG	0.619584
ORtg	0.251895
DRtg	-0.265725
OWS	0.598172
DWS	0.626173
WS	0.677791

Correlation heat map of variables:

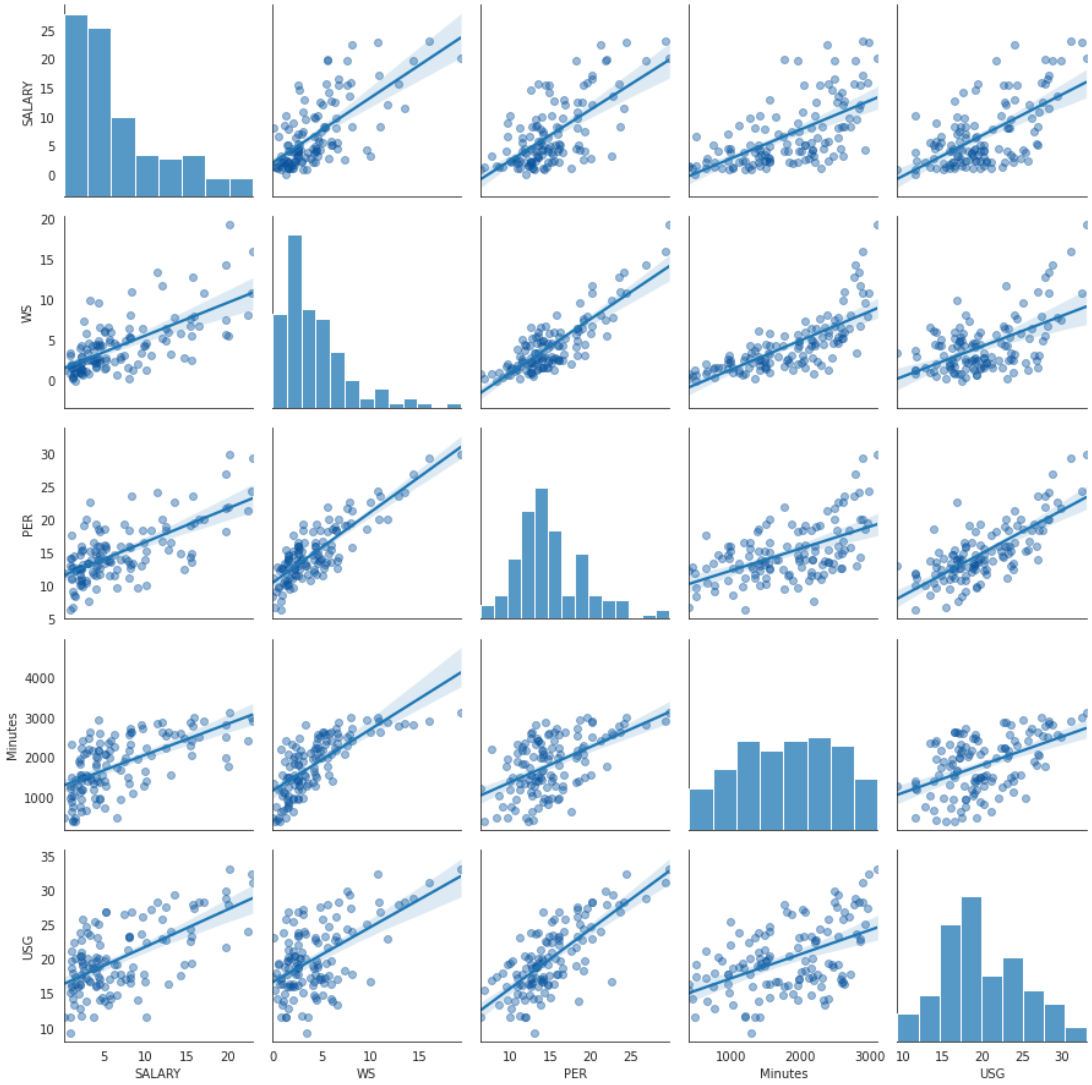


Based on correlation matrix that we have Minutes, PER, USG, OWS, DWS and WS have correlation bigger or equal to 0.5. This chosen variable will be used in our KNN model and Random Forest regression model.

We did statistical summary for Minutes, PER, USG, OWS, DWS and WS:

Statistical summary of other variables				
Variable	Minutes	PER	USG	WS
Count	126	126	126	126
Mean	1809.0159	14.9643		4.1825
Standard Deviation	703.5615	4.3063	4.9621	3.4027
Minimum	393.0000	6.3000	9.3000	-0.1000
25%	1290.2500	12.4250	16.4250	1.7250
50%	1859.0000	14.0000	19.1000	3.1500
75%	2405.0000	16.7500	23.2750	5.5750
Maximum	3122.0000	29.8000	33.0000	19.2000

Scatterplot for Minutes, PER, USG, OWS, DWS and WS:



Linear regression model:

For our initial model, we removed OWS and DWS as they have potential collinearity with WS. ORB and TRB are removed as well as they have potential collinearity with DRB. Furthermore, we remove TRAIN_ID from our variable as TRAIN_ID is a unique identifier of each players, and SALARY as its used as our response variable. This leave us with 15 variables to be tested.

Methodology and Analysis:

We chose variables based on hypothesis testing, in the hypothesis testing we have null hypothesis: predictors equal to zero and alternate hypothesis: predictors are not equal to zero, if we have p-value lower or equal with 0.05 we reject the null hypothesis and p-value larger than 0.05 we retain the null hypothesis. The closer the p-value to 1 the more evidence to retain the null hypothesis and the closer the p-value to 0 the more evidence to reject null hypothesis.

Fit the initial predictors to the model, we detect “Teams” has the largest p-value (0.989), we remove “Teams” from the list of predictors. Re-fit the new list of predictors (predictors without “Teams”), we detect “PER” has the largest p-value (0.945), we remove “PER” from the list of predictors and we re-fit the new list of predictors. We continue to do this until all variables that fitted into the model has

p-value less than 0.05, we removed “DRB”(0.826), ”AST”(0.768), “STL”(0.600), “POSITION”(0.565), “BLK”(0.855), “TOV”(0.660), “Age”(0.503), “TS”(0.396) and “ORtg”(0.691). We have our final model, shown in the table below:

Adjusted R-squared	0.628
R-squared	0.643
AIC	675.0
BIC	692.0
SER	3.99156

Predictor	**Coefficient**	**Standard error**	**t**	**p-value**	**95 CI lower**	**95 CI upper**
Intercept	38.1388	10.827	3.523	0.001	16.072	59.575
Games	-0.1807	0.041	-4.446	0.000	-0.261	-0.100
Minutes	0.0048	0.001	5.262	0.000	0.003	0.007
USG	0.2837	0.079	3.573	0.001	0.126	0.441
DRtg	-0.3237	0.096	-3.357	0.001	-0.515	-0.133
WS	0.3266	0.161	2.025	0.045	0.007	0.646

Based on the final model we have, our AIC value was reduced from 718.0 in the initial model to 675.0 in the final model, was reduced from 854.1 in the initial model to 692.0 in the final model. The higher BIC value than AIC value is expected as BIC gave heavier penalty on model complexity or model with more predictors/variables has higher BIC value compared to simpler model/model with less predictors.

According to the summary of the final model, the estimated relationship can be illustrated as following.

$$\widehat{\text{Salary}} = 38.1388 + 0.3266\text{WS} + 0.2837\text{USG} - 0.1807\text{Games} + 0.0048\text{Minutes} - 0.3237\text{DRtg}$$

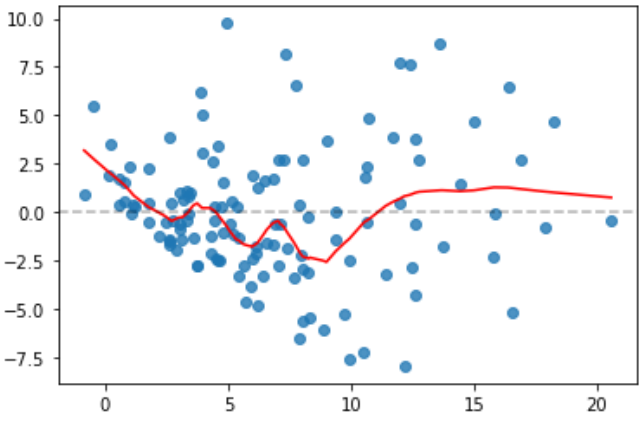
This estimated relationship provided the following information:

- As other variables remain constant, 1 unit increase in win share will increase the NBA player salary by 0.3266 millions.
- As other variables remain constant, 1 unit increase in usage percentage will increase the NBA player salary by 0.2837 millions.
- As other variables remain constant, 1 unit increase in games played will decrease the NBA player salary by 0.1807 millions.
- As other variables remain constant, 1 unit increase in minutes played will increase the NBA player salary by 0.0048 millions.
- As other variables remain constant, 1 unit increase in defensive rating will decrease the NBA player salary by 0.3237 millions.

	Test RMSE	**Test adjusted R^2**	**Test R^2**	**MAE**
OLS	3.985147	0.470083	0.474289	2.98299

OLS Assumptions:

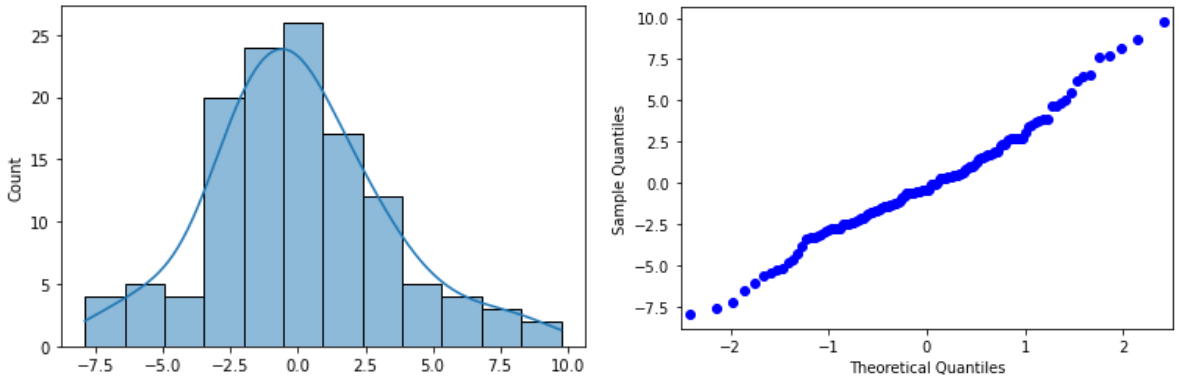
1. **Linearity**: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + B_5 X_5 + \varepsilon$ is the true population model.



From observing the LOWESS line of the residual against the fitted values, there are nonlinearity detected, as the model under-predicting salary for 0 < salary < 4 million salary, over-predicting salary for 4 < salary < 12 million and under-predicting salary again for 12 < salary < 22 million. The line is not flat (there is a significant up and down trend along the line), we could conclude that linearity assumption is not met. This problem can be overcome by doing variable transformation to capture this nonlinearity.

2. **Exogeneity**: $E(\varepsilon|X) = 0$.

Statistical summary of sample residuals	
Min	-7.9141
Max	9.7554
Mean	2.6434 * e-15
Variance	11.3836
Skewness	0.3675
Kurtosis	0.4368



From the statistical summary of the sample residuals, the distribution of sample residuals is a bit right skewed and there is some outlier detected due the skewness of 0.3675 and the kurtosis of 0.4368. Furthermore, as suggested by the sample mean of residual of 2.6434 * e-15 which is approximately 0 and the fact that the LOWESS line deviate much from zero, it is reasonable to conclude that exogeneity assumption is not met despite the QQ-plot show the residual are normally distributed and as mentioned before the distribution of the residuals are skewed to the right. This problem can be overcome as well by doing variable transformation.

3. **Independence**: The data pairs Y_i, X_i are i.i.d.

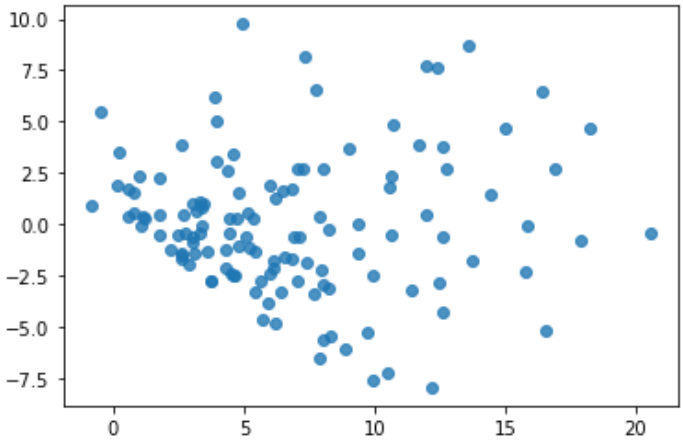
Without any further description of data collecting procedure, the validity of this assumption in this data set cannot be assured. Nevertheless, this assumption may assume to be true for NBA player salary unless a player plays for two team or players lied about their salary. This should also be true for the predictors used in the model unless there is a measurement error when collecting the data. Hence, the assumption of independence and identical distribution of data may be satisfied for this data set.

4. **4th moment exists**: $E(Y^4), E(X^4)$ are finite $< \infty$.

Considering the salary of NBA player, this can reasonable be a bounded variable since salary cannot be lower than 0 and should have a certain maximum point logically. Hence, the fourth moment of NBA player salary can be regarded as a finite one.

For the predictors, all of them are reasonably bounded variable, they have a certain maximum and minimum point. We can conclude that the fourth moment assumptions are met for both response variable and predictors variable.

5. **Constant error variance**: $Var(\varepsilon|X) = \sigma^2$.



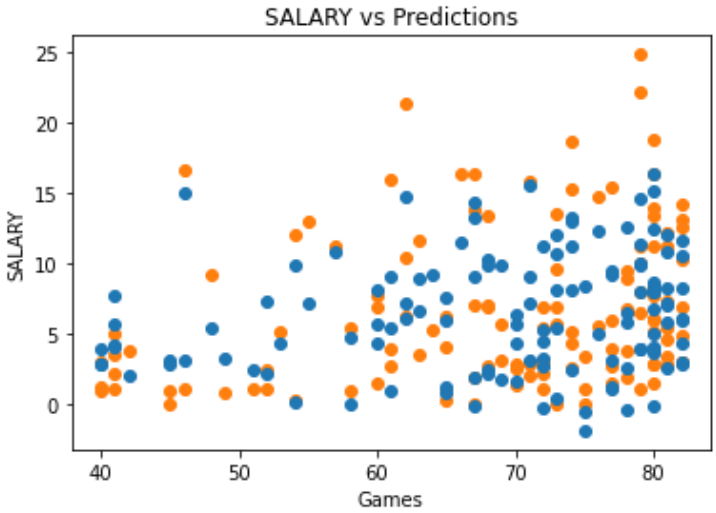
By observing the scatter plot of residuals against fitted values, there is a clear fanning out pattern when the fitted value is increasing. This fanning pattern show sign of heteroskedasticity. Hence, the assumption of constant error variance is not met. This can be overcome as well by variable transformation. We could use "Heteroskedastic-robust" covariance when fitting the models, this allows us to relax the constant variance error assumptions.

6. **No perfect multicollinearity**.

Variable	Games	Minutes	USG	DRtg	WS
VIF	2.27498	4.40801	1.63610	1.32752	3.17570

While the VIFs for USG and DRtg are lower than 2 as indicated above, while VIFs for Minutes, WS and Games higher than 2. However, all figures observed are still lower than 5, so there may not be a fair degree of variance inflation and collinearity in the data. Furthermore, the average VIF is below 3 (2.5 6446) so, again, collinearity should not be a huge issue. Hence, the assumption of no perfect colline arity is satisfied

We can see accuracy of our model represented by the plot below, our fitted value (Blue) with the observed value (Orange) based on "Games" variable and response variable "SALARY":



Disadvantages of the models:

Sensitive to outliers: As the fourth moment assumptions need to be satisfied in the model, MLR with least square the model accuracy will be decreased, this could be overcome by using least absolute deviation model rather than the least square model, as LAD use median instead of mean.

Sensitive to non-linear variable: MLR first assumptions is about linearity, so if predictors variable doesn't have any linear relationship with the response variable, we need either to do variable transformation or we leave out the variable from our MLR model. But the variable transformation suffers from bias when we change them back from their "transformed" form into their "original" form.

Multicollinearity: MLR model are prompt to multicollinearity issues since multicollinearity wear away the statistical significance of an independent variable in the model.

K-nearest neighbor model:

We only used predictors that correlated larger or equal to 0.5 with our response variable ("SALARY"). From the IDA above initially we have Minutes, PER, USG, OWS, DWS and WS but due to potential multicollinearity we decide to take "WS" only and remove "OWS" and "DWS". the main reason we only use variables that has correlation larger or equal with 0.5 is due to avoid curse of dimensionality, which indicated by an increases of test error or the test error remain the same when we add a new variable, this will cause the predictions to get less and less accurate.

Methodology and Analysis:

We created knn_test function which used to find the best number of K neighbours in the range of 1 to 50 for the inputted predictors. We fit the KNN model, our KNN model use an iterated K neighbours, our train predictors and train response data. Furthermore, the model that we fitted before are evaluated using cross-validation method, we set 10 folds in our CV. We repeat this procedure in each iteration with different K neighbours. We are finding model with smallest CV RMSE to find the best K neighbours.

When we find our best K neighbours with the smallest CV RMSE, we refit the model with test predictors and used this as our "predictions" to compared it with our test "SALARY". We are finding the smallest test RMSE for best model selection.

First, we select "WS" to fit the model as "WS" has the highest correlation compared to the other selected variable. We calculate the CV RMSE and test RMSE to evaluate the performance. Then we continue to add "PER" as our second highest correlation predictors, when we add "PER" to our model the test RMSE increase instead of decreasing this is an indication of curse of dimensionality. We drop "PER". Then we add "Minutes" as our third variable, the test RMSE decreased, so we keep "Minutes" in our model. Adding "USG" as our fourth or final variable, test RMSE of the model increases. We observed and conclude that the model with two predictors ("WS"," Minutes") with 22 K neighbours as our best model. The simulation mentioned above shown by the table below:

Variables	Test RMSE	CV RMSE	K-neighbours
"WS"	4.2088	4.2900	22
"WS", "PER"	4.4035	4.3518	22
"WS", "Minutes"	4.1955	4.2531	24
"WS", "Minutes", "USG"	4.6920	3.9033	7

We could find the Test RMSE, CV RMSE, Test R^2 and MAE of our best model with two predictors ("WS", "Minutes") and 22 K neighbours:

	Test RMSE	CV RMSE	Test R^2	MAE
KNN	4.1955	4.2531	0.4099	3.1911

Our test RMSE is 4.1955 which is close to our RMSE target (4.1 millions) given in the task. The test RMSE indicates how spread out the residuals compared to our fitted value. Based on this result we could expect that our predicted salary/fitted salary to be be ± 4.1955 millions from the observed salary.

Assumption of K Nearest Neighbours:

K nearest neighbour is a non-parametric model, there is no formal assumptions used in the model. This model can handle skewed distributed data or any distribution of the data (we usually used non-parametric model when the distribution of data is not normally distributed). This model can handle both numerical and categorical data. The only assumption this model needed is independence and randomness. In this case, without any further description of data collecting procedure, we can argue that all assumptions are met.

Disadvantages of the models:

Sensitive to irrelevant feature: This could cause our model to overfit the data set and make a poor prediction in the future data set.

Curse of dimensionality: This is the main issue when building a KNN model, our test RMSE is increased as our model complexity increase or numbers of our predictors increase. This cause our KNN model to have poor accuracy and eventually break down.

Computationally expensive: For a large data set, as the computing cost of an existing point with a new point are huge and this will slow down the computation process.

Random Forest Regression:

In a random forest regression, we create many decision trees, decision tree determines their outcome based on multiple questions and conditions forming "branches" in the tree. Each decision tree learns from different sub samples of observations and different combination of variables. We constructed our random forest by choosing the number of decision trees to grow and number of predictors (question) to consider in each tree. In a random forest tree, we randomly select rows in our data frame with replacement, replacement is needed so we don't end up with the same set of rows in each decision tree and we randomly select the appropriate number of predictors from the data frame. We repeat this procedure many times, prediction is made by taking the majority outcome from all decision trees in the "forest".

Methodology and Analysis:

For our random forest tree, we use same initial variables as we used in our KNN model, predictors with correlation coefficient larger or equal to 0.5. we iterate from 1 to 50 N estimators. We fit each iterated N estimator, train response and train predictors and to allow us to have a reproducible outcome, set random seed as 1 in the Random forest regression model. Furthermore, we use cross validation to evaluate the model, we set our CV's fold to be 10 folds for each model. We repeat this procedure in each iteration with different N estimators. We are finding model with smallest CV RMSE to find the best N estimators or the best numbers of tree to grow.

When we find our best N estimators with the smallest CV RMSE, we refit the model with test predictors and used this as our "predictions" to compared it with our test "SALARY". We are finding the smallest test RMSE for best model selection.

First, we select "WS" to fit the model as "WS" has the highest correlation compared to the other selected variable. We calculate the CV RMSE and test RMSE to evaluate the performance. Then we continue to add "PER" as our second highest correlation predictors, we could see the Test RMSE and CV RMSE decreases. Then we add "Minutes" and "USG" as our third and fourth variable respectively. We observed the test RMSE and CV RMSE keep decreasing after we add more variables and conclude that the model with four predictors ("WS","PER","Minutes","USG") with N estimator as our best model. The simulation mentioned above shown by the table below:

Variables	Test RMSE	CV RMSE	K-neighbours
"WS"	4.9285	4.7574	20
"WS", "PER"	4.9608	4.4520	13
"WS", "PER", "Minutes"	4.7504	4.2013	11
"WS", "PER", "Minutes", "USG"	4.5953	4.1893	13

We could find the Test RMSE, CV RMSE, Test R^2 and MAE of our best model with four predictors ("WS","PER","Minutes","USG") and 13 N estimators:

	Test RMSE	CV RMSE	Test R^2	MAE
KNN	4.5953	4.1894	0.2921	3.5192

Our test RMSE is 4.5953 which is close to our RMSE target (4.1 millions) given in the task. The test RMSE indicates how spread out the residuals compared to our fitted value. Based on this result we could expect that our predicted salary/fitted salary to be be ± 4.5953 millions from the observed salary.

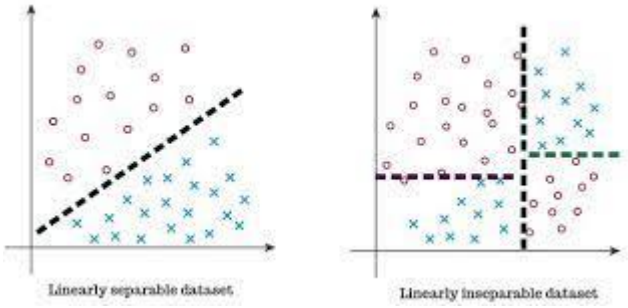
Assumption of Random Forest Regression:

Random forest regression a non-parametric model, there is no formal assumptions used in the model. This model can handle skewed distributed data or any distribution of the data (we usually used non-parametric model when the distribution of data is not normally distributed). This model can handle both numerical and categorical data. The only assumption this model needed is independence and randomness. The only assumption this model needed is independence and randomness. In this case, without any further description of data collecting procedure, we can argue that all assumptions are met.

Disadvantages of the models:

Overfitting: Decision trees get very complex easily, without giving a penalty on complexity, the model will keep adding more variables, this could cause severe overfitting to the model. The model will have poor predictions for future relevant data.

Unflexible model: Decision tree only could make decision parallel to either x or y axis, they can't make decision diagonally.



Conclusion of Task A:

We conclude that our best model is our K nearest neighbour model as all assumptions are met in the model and it has lower test RMSE compared to our Random forest regression model, despite our Multiple linear regression provide the lowest test RMSE, some assumptions of MLR are violated and further improvement in the model are needed for us to meet all the assumption.

TASK B

Question 1:

a) Observed data need to be partitioned into training and validation set is important, as we need training data to train our model, from the train data as well our algorithm/model learn the relationship between response and predictors variables. Usually in a supervised machine learning, our training data is labelled with known outcomes. Validation sets is used to evaluate how accurate our model to identify relationship between our predicted outcomes

that we gain from fitting our model with new predictor's value from the validation sets and the known outcomes in the validation sets. By partitioning our data, this helps to develop a highly accurate model to predict on relevant future data, this will lead to better decision, greater confidences, and model accuracy when it given a new relevant data.

b) Random partition is used to ensure that each data in the population have an equal chance to be part of a partition. This randomization process allows us to generalize the result we obtain from one of the data partitions into larger population of the data. For examples we collect data of equal number of child respondent and equal number of adult respondent, and we are going to partition the data into 80% and 20%, with random partition it will ensure there is equal number of adult and children in both 80% partition and 20% partition, random partition also prevent either number of adult or children to dominate in either 80% and 20% partition.

Question 2:

a) The initial graph "Fertility" vs "PPgdp" suggest that there is no clear pattern of linearly relationship between two variables. This result of the assumptions of linearity/ first assumption to be violated, when we draw the LOWESS line there could be data that being uncaptured by the line due to non-linear relationship of the data, this result in poor accuracy of the model. Converting both variable to log scale gave us this equation:

$$\log(Fertility) = \beta_0 + \beta_1 \log(PPgdp)$$

This equation suggested that the change A 1% change in PPgdp is associated with a $\beta_1\%$ expected change in Fertility. The limitation of log-log model is a potential bias when turning our log value back to its original value. Furthermore, this bias might cause misleading conclusion of our data.

b) We need to remove Purban based on variables selection using hypothesis testing method, as the p-value of Purban is larger than 0.05 this indicate us to retain our null hypothesis which mean Purban's predictors is equal to zero. The p-value of Purban is close with our alpha indicate weaker evidence to reject the null hypothesis. Purban has high correlation coefficient of 0.78 with log(PPgdp) as well, this indicate multicollinearity could occurs between two variable in the model. Inconclusion despite the weak evidence to conclude that the Purban's predictor is equal to 0 or insignificant in the model, there could be potential multicollinearity between Purban and log(PPgdp), so we conclude to remove Purban from the model. Our final model would be:

$$\log(Fertility) = \beta_0 + \beta_1 \log(PPgdp)$$

References:

Tarr, G (2020). DATA2002 Data Analytics: Learning from Data, Lecture 30 Decision Trees and Random Forest Regression. University of Sydney, Sydney Australia.

Jajo, N (2020). QBUS2810 Statistical modelling, Lecture MLR assumption. University of Sydney, Sydney Australia

Chatterjee, M (2020). A Quick Introduction to KNN Algorithm.
https://www.mygreatlearning.com/blog/knn-algorithm-introduction/

Azika, A (2019). Decision tree: Part ½.

https://towardsdatascience.com/decision-tree-overview-with-no-maths-66b256281e2b

Data Robot. What are Training, Validation, and Holdout?
https://www.datarobot.com/wiki/training-validation-holdout/

Springer (1997). The problem of multicollinearity. Boston,MA

https://link.springer.com/chapter/10.1007%2F978-0-585-25657-3_37

Matange, Y (2021). Who are the highest-paid NBA players for the 2021-22 season?

https://ca.nba.com/news/who-are-the-highest-paid-nba-players-for-the-2021-22-season-stephen-curry/1dduhjmusvdvn18iikepy82d5g