1 People Analytics (GoT...Analytics) (30 Pts)

The CEO is concerned regarding recent events at the company, and wants to initiate a new

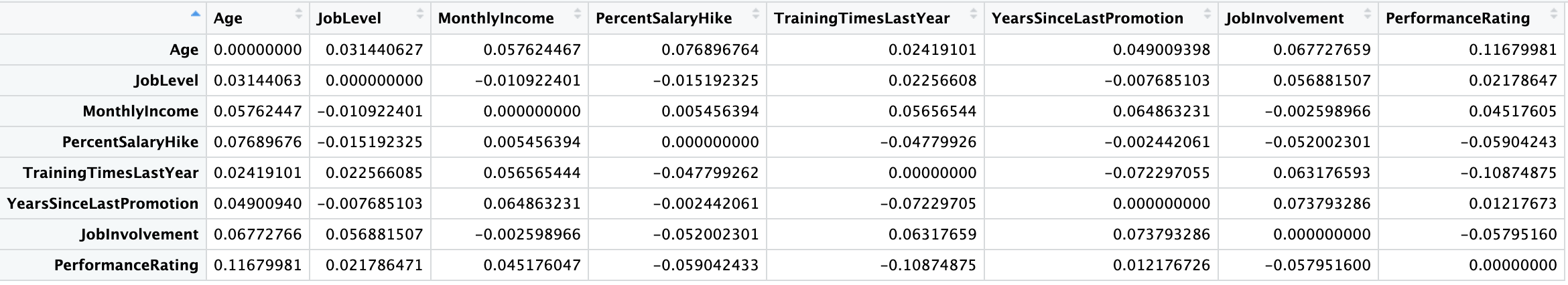
program. It is concerned with gender discrimination and equitability across the board. They are seeking to also increase overall job satisfaction. Using the employee datasets (from the midterm), provide a determination if there is gender discrimination in the workplace, and provide an assessment of factors that should be looked at for job satisfaction and how these factors can be used. This should be a very comprehensive review.

## **gender discrimination and equitability:**

To define what’s “gender discrimination and equitability”, we can say two genders have different numbers of **“Income”, ”SalaryHike”, “promotion”, “Joblevel”**and **“Jobinvolvement”**.

First, I split the dataset to two based on gender. To check about the difference, we have to check the correlation difference between each variable.

Correlation difference:



In this data, we can say that females have more advantage in **age**, **TrainingtimeLastYear**. This means if a female is older or had more training last year, her salary and salary hike will be higher than the man with the same condition.

On the other hand, Men have more advantages in**PercentSalaryHike** **and Job involvement.**This means a male has more chance to get a higher salary hike even if he does the same thing as another female.

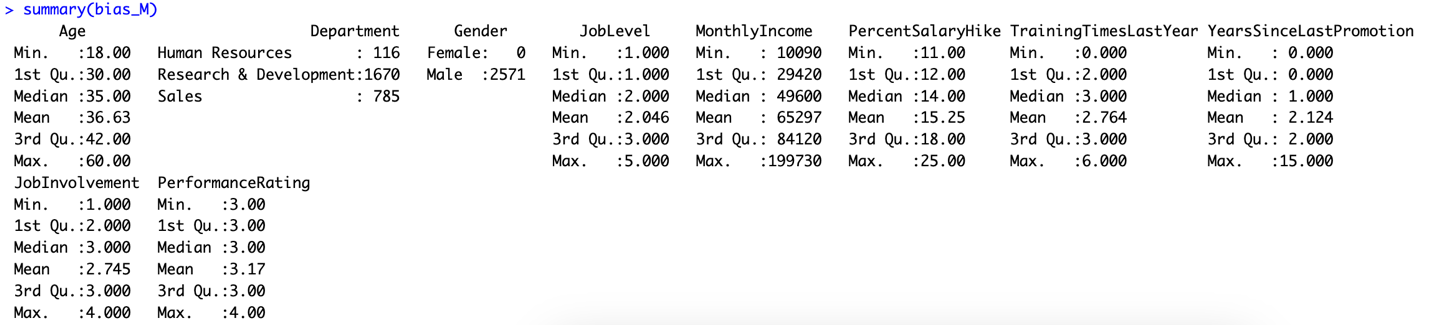
However, all the differences are not significant. We can not say this company has**gender discrimination and equitability.**

Second, we have to check about the Statistical data to find if those distribution is bias or not.

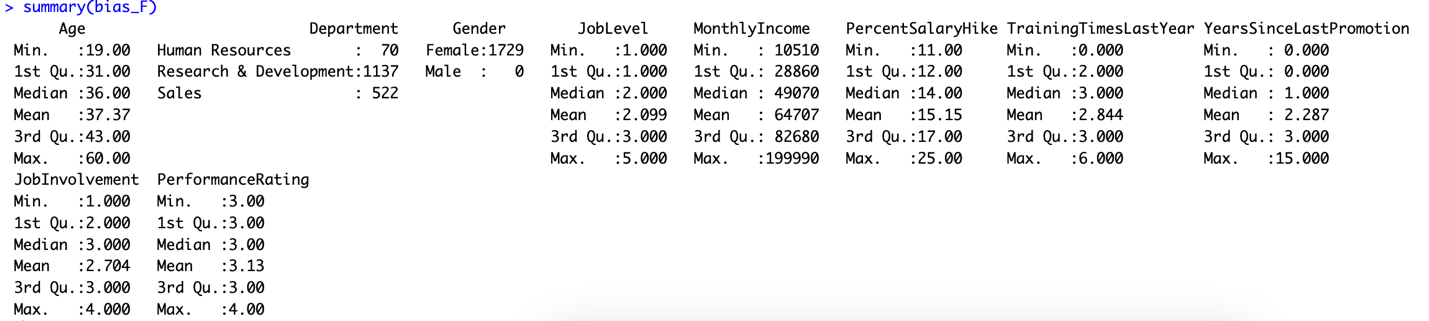
We focus on **“Income”, ”SalaryHike”, “promotion”, “Joblevel”** and **“Jobinvolvement”.**

However, those value are significant different, we can not say there’s a **gender discrimination and equitability** in this company.

Male :



Female :



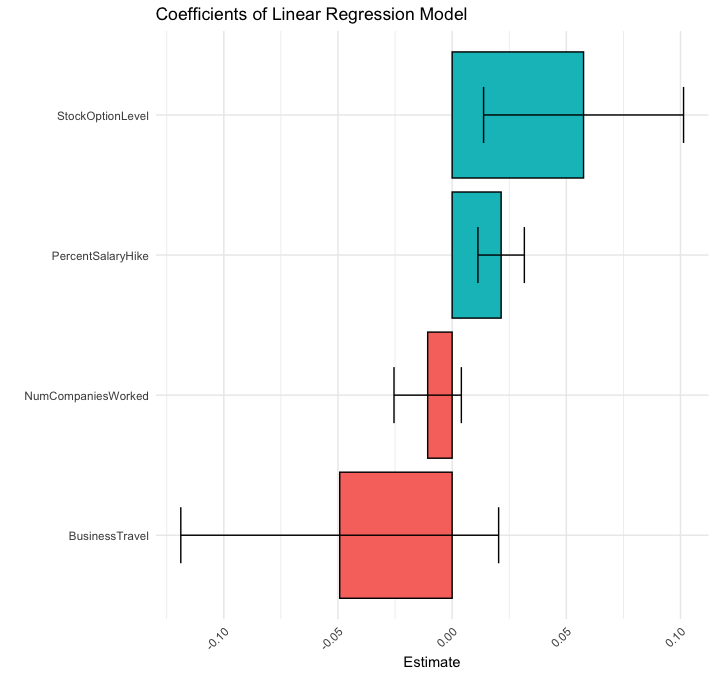
**increase overall job satisfaction :**

For this investigation, we compare all the combinations to determine what variables are highly related to Job satisfaction by linear regression algorithm.

Finally, in this chart, we can say that “**StockoptinLevel”, “PercentSalaryHike”, “NumCompaniesWorked”,** and **“BusinessTravel”** are the most influential variables.

According to data, if we increase Stockoptionlevel and percent salary hike, the employees will have more job satisfaction. On the other hand, if we increase the business travel rate, employees will have less job satisfaction.

Additionally, the higher the number of companies that employees worked before, the lower their job satisfaction.



2 Handwriting Recognition: ENDGAME (30 Pts)

1. Why was your colleague wrong in the first place?

There’s a missing value for the target variable, which will affect prediction significantly. Because the neural network algorithm is a Supervised method algorithm, it cannot recognize itself if we do not import specific target variables. And colleague did not import “6” for this algorithm’s target variables. Therefore, when predicting new dataset which include “6” for the actual value, the prediction will be an error.

1. Your colleague reran the code and claims that a one layer ,40 hidden node model actually works best, since there are ten numbers and each of the nodes gives a 25% probability of hitting a number from 0-10, and thus each hidden node can be explained clearly. Explain why this is a good rationale or not.

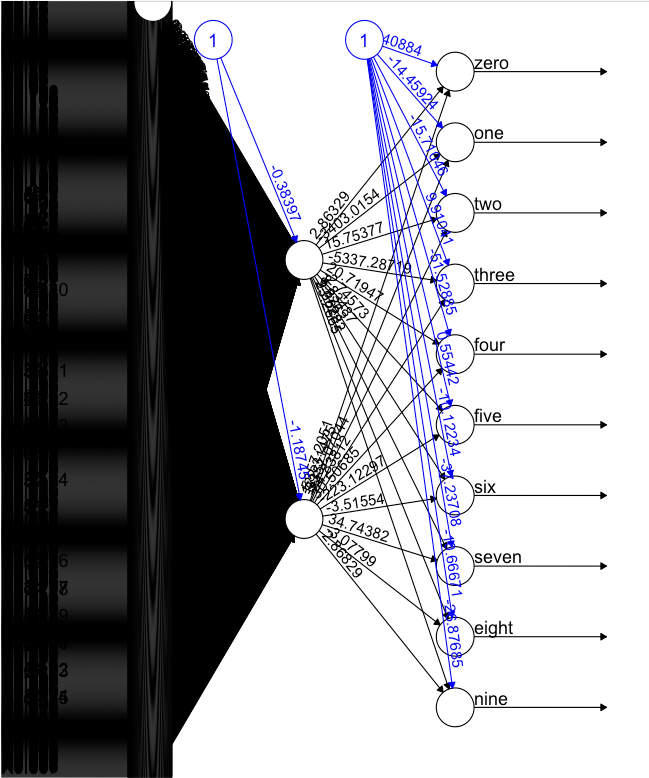
Indeed, higher nodes will have higher accuracy. We should input as many as we can. However, we also need to concern about overfitting. When a neural network has too many neurons, the model begins to acquire the noise of the training data rather than the general trends. This is referred to as overfitting. Overfitting may cause the model weights and bias values to become very large, decreasing the accuracy.

To avoid overfitting, we should use this equation to set the upper limit of nodes :

In this case, the result was 2. It will be overfitting if we set nodes above 2.

(c) Re run your neural network, however, this time you have a training set of 60,000 and the

test set of 10,000 (this may take some time to run).



Observation 1000

“accuracy : 0.632"

Observation 1500

accuracy : 0.816666666666667"

time spend: 1368 second

(d) Consider all the techniques we learned in class

• Can any other technique be used to perform the classification. If so list the techniques with a one-line rationale as to why

K-Nearest Neighbors can be used to perform in this case too.

However, k-nn is not appropriate for using neural networks. The KNN algorithm is extremely sensitive to outliers due to its distance-based. If there are many outliers, these data will impact the classification or regression results of the algorithm for the test samples.

The dataset is clean In this case, so it does not affect that much.

• Run the techniques in part 1, and provide the final answer for that technique along

with the training error rate and test rate (be sure to indicate any other trials you

had).

k = 10

accuracy = 0

for (i in 1:nrow(glm\_df\_test)) {

pred <- knn(glm\_df\_train[,-785], glm\_df\_test[i, -785], class, k)

if (pred == glm\_df\_test[i,785]) {

accuracy = accuracy + 1

}

}

accuracy <- accuracy/nrow(glm\_df\_test[,-785])

print(paste0("Accuracy: ", round(accuracy, 3)))

We use KNN to classify the training set, and use this model to predict the testing set.

The Accuracy is 83%.

(e) Consider everything above, write a summarized conclusion of everything above you would hand to a Chief Analytics Officer.

First, the machine learning algorithm is a highly sensitive data analysis method. We have to take care carefully of every step we take. Otherwise, those mistakes will bias our analysis and produce some wrong that is different from what we wish. Colleague has two errors in this case.

One is missing value in target variables, significantly affecting the final result. If new data include this missing value imported to this model, the model can never predict this missing value right. Another one is miss estimate the nodes because the neural network’s highly sensitive if the analyst only takes care of prediction accuracy and forget the overfitting problem. The overfitting result may have great accuracy on the train set, but when we test the new data, it will be unstable and have lousy accuracy.

(f) After this, what should happen to your colleague?

He should learn and quit, because I don’t want to write this type of report again!

3 Performance Optimization (Return of the FIFA) (30 Pts)

The FIFA Director of Analytics isn’t completely convinced with your prior analysis. The belief is that your analysis isn’t useful. Use the original fifa.csv dataset.

1. Explain how your prior FIFA analysis was useful and how it can be used.

For FIFA analysis, we used Factor analysis to make Dimension reduction.

Factor analysis is a statistical technique used to decrease the number of variables in data by identifying and grouping them into smaller factors. This helps simplify the processing and interpretation of the data and extract the most crucial information we should focus on when we analyze the player. Otherwise, we have to check 89 variables when we evaluate every player.

Ultimately there needs to be a unifying factor across all techniques used. If the primar3y

question is how to build a team then your job will be to develop a methodology that will

assist in this, therefore:

1. How would you group players ? Be sure to establish what fields are necessary prior to grouping and indicate how you grouped them. Then describe the groups.

We can use the k-means cluster to group players. First, we choose **"Age," "Overall," "Potential," "Value," "International.Reputation" "Weak.Foot", "Skill.Moves", "Height," "Weight," "LS," "LCM," "LWB," "Crossing," "Finishing," "HeadingAccuracy," "FKAccuracy," "LongPassing," "Acceleration"**those variables that are based on our factor analysis.

As a result, Group 1 would be **a high-value player**;

Group 2 would be **younger players**;

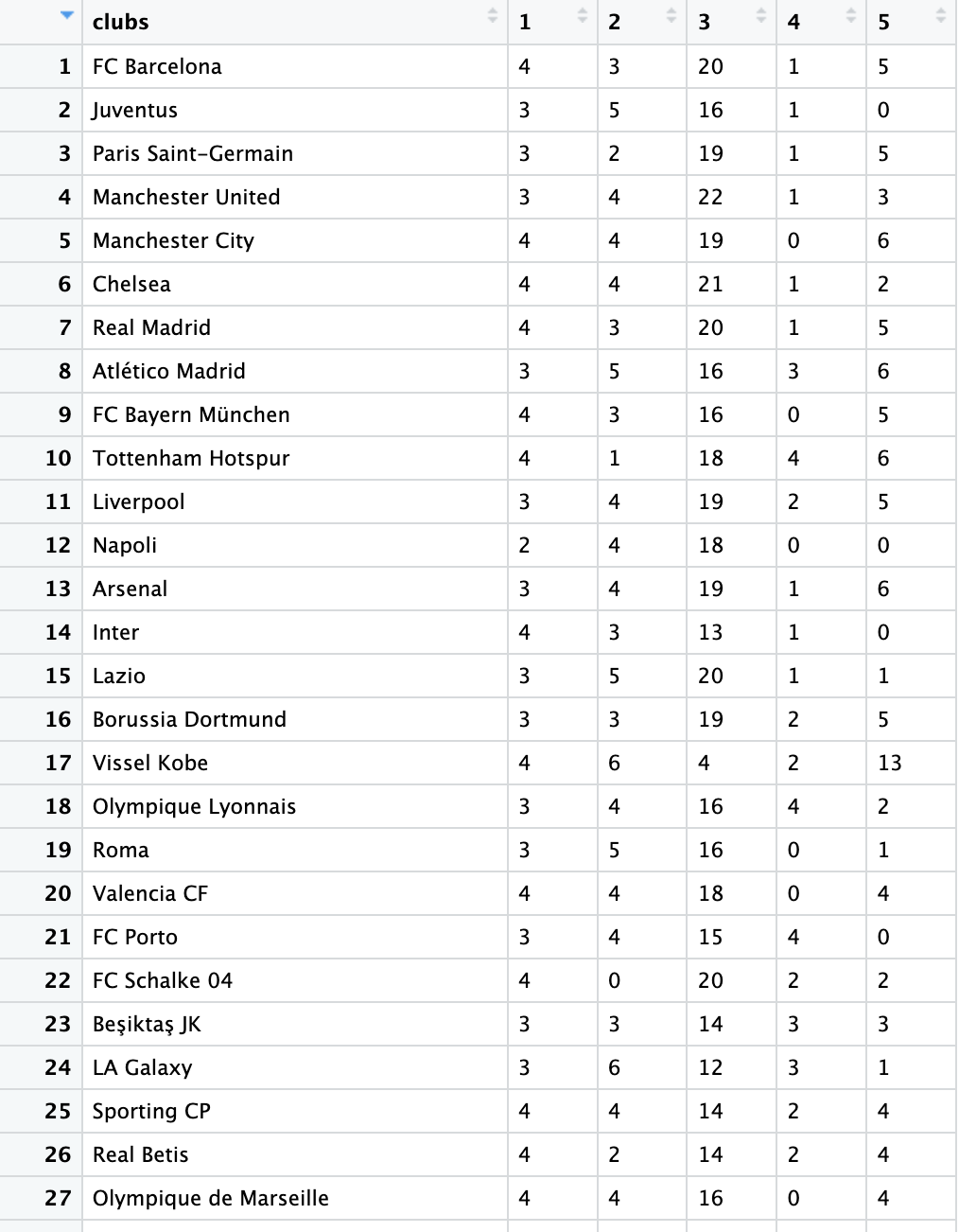
Group 3 would be **a mid-level player**;

Group 4 would be **a high-potential player**;

Group 5 would be **weak-level players**;

1. Using the groups can you assess the clubs. Explain if this is useful.

By using the group, we can check each club’s player-level distribution. For example, Vissel Kobe have 13 high potential player(cluster 4), which means their club has a lot of great players.

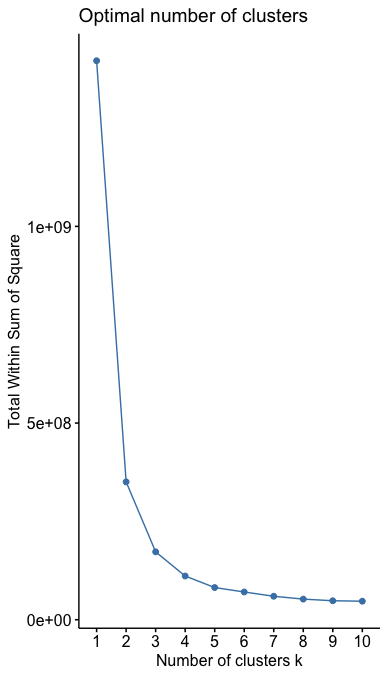
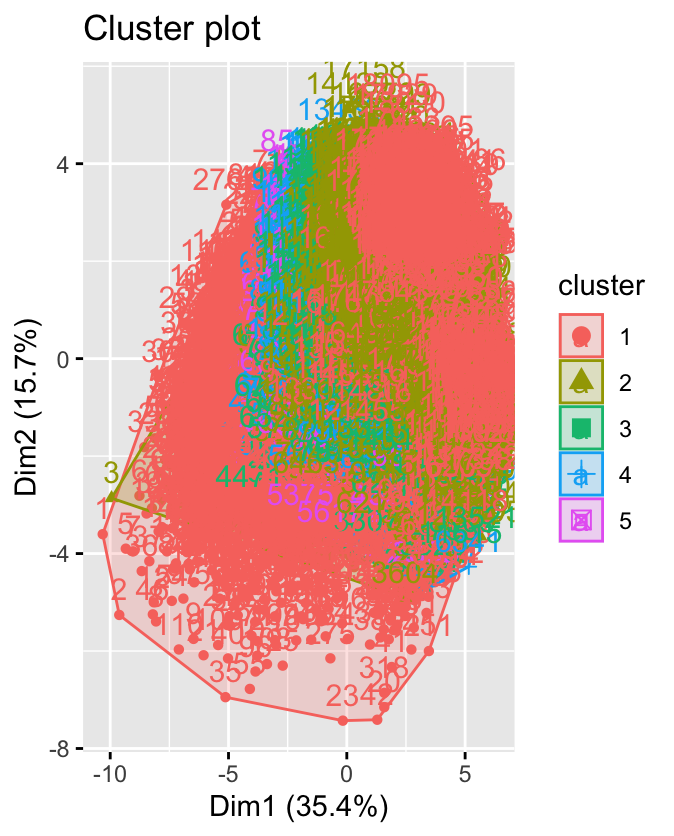


(c) How can your initial analysis (from midter) be integrated with the previous two bullet

points?

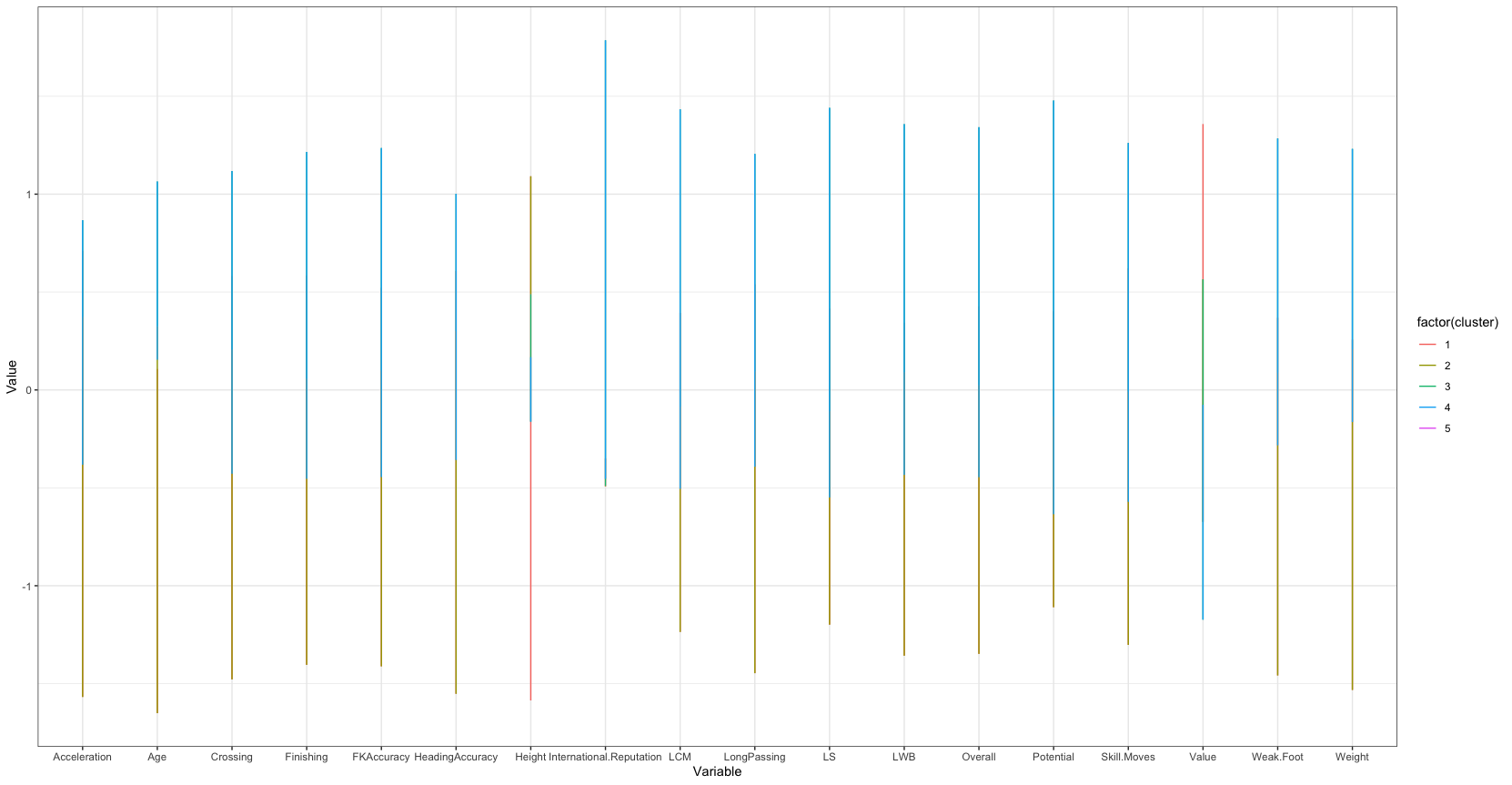
For factor analysis, we can analyze which variable is helpful in this case. Also, factor analysis can reveal commonalities and correlations among observed variables and help us understand the internal relationships and structures among variables.

After factor analysis, we will only have variables that are filtered, which means the variables we determined it is impactful in this case. Furthermore, We can use hierarchical clusters to group the data by the mathematical method. Hierarchical cluster can classify and generalize data by dividing samples into different clustering groups. This means it will mark each player as a group number, which will make the analysis of every player or club easier.

For example, we can visualize the distribution of each trait (variables).

By this chart, we can say group 4 (high potential player) has advantage in most trait.



(d) Can you find any other useful and interesting analytics (from our techniques) that

would be helpful to the Director of Analytics.

In this dataset, we also can use ANOVA (Analysis of Variance) to measure if there are significant differences in means between the groups. For example, using ANOVA, we can observe the difference between each club.

Furthermore, ANOVA can also be used to analyze the effect of interactions between different factors on the observed variables. This helps us to know if the variable would affect another variable and how effective it is.

4 Question 4 - Theoretical Questions (10 Pts)

Provide no more than a paragraph for each, be concise.

(a) Michele Piccinno a researcher in Artificial Neural Networks stated (2016), “Neural Networks are the second best way to solve a problem. The best way is to understand the

problem”. Provide an opinion based on your knowledge of whether you agree or disagree

with this statement?

I agree with this statement.

First, Neural networks can handle non-linear relationships and are suitable for large-scale datasets. However, it is less interpretable because of Black box properties; they do not easily explain how the model works internally and the decision process.

Second, Neural networks need proper hyperparameter tuning. Otherwise, they may be overfitted or underfitted. For example, when deciding on neurons and hidden layers, adding one node can change the result significantly.

(b) Explain how accuracy is measured across the techniques and why is it appropriate forthose

techniques?

We could use K-fold cross-validation to measure the accuracy. In the K-Fold method, we can cut the training data into training and test sets in order, and the test set in the K-Fold can be used as the validation set. we freely control k, and in each iteration, one set is selected as the validation set, and the remaining (k-1) sets are used as the training set.

This method can reduce the bias of the model training on the data set. In addition, the training and testing sets are fully utilized and learned.

(c) What are some of the challenges with data when starting a project and what would your

initial steps be?

In my opinion, the most important and complex step is data cleaning; it also is the first step when we handle the data if collecting the data is not the first step.

The initial step would be to decide which variables should be removed. We can filter

the variables that do not relate to the target variable subjectively, then remove the variables with missing values or error-measured data. Furthermore, define what is outliner and remove them. Finally, combine or remove the highly correlated variables. However, this step depends on the algorithm. Some algorithms are sensitive to multilinearity problems, but others are not.

(d) Explain how a decision tree can be used to predict a continuous variable. How are the

results interpreted?

Regression Trees is a model for predicting a continuous variable. We use recursive partitioning techniques that rely on various predictors to partition the data. The prediction is then made by calculating the average response values for all observations within that particular subgroup.

(e) Provide an example of an ethical problem with data mining, and what an analyst should

consider and how they should mitigate the problem, if possible.

All the biological problems that have been solved by data mining will have an ethical issue.

Because the machine learning prediction is based on a training set, if the training set is not precise or biased, the result may be biased to specific things or races.

Furthermore, we need to determine if the model is 100% accurate. Therefore, when dealing with life issues, such as the death penalty or social credit control, we should set up a final step that is evaluated by a professional person.