FIFA 2019, Exploring Players Statistics and Build Your Dream Team
Group Member: Suzy Gao, Yufan Luo, Zhijun Liu, Yu Yang

Write Up

1. Statement of the problem

"FIFA19", released on September 18th, 2018, is a football simulation video game developed by EA,. In "FIFA19", players hire characters (football players) from their original clubs and then build their own teams for all season games. Since the features and images of characters are created based on the data of real-world football players, this game not only attracts football fans, but also serves a reference when clubs hire players and make up their teams.

For this FIFA dataset, we want to do the following tasks:

- High level exploratory data analysis.
- Build a predictive model to predict possible position for a player based on the skill sets of that player.
- Build a replacement system. Through the system, users can find a similar replacement for a particular player, and the resulting output displaying a list of players and their attributes who are most similar.

2. Techniques, Process and Results

2.1 Data cleaning

- Drop the useless variables. In this case, we dropped variables, such as 'X', 'Photo', 'Flag', 'Club.Logo', and 'Loaned.From', which are meaningless.
- Convert Value, Wage and Release.clause into numeric variables, and set the default unit as 1000€. For example, we turned €110.5M into 110500(K€).
- Split Work.Rate into attack_work_rate and defend_work_rate. For example, we split Work.Rate ('Medium/ Medium') into attack_work_rate ('Medium') and defend_work_rate ('Medium').
- Recalculate the height and set the default unit as cm.
- Change the data type of Joined into date type. For example, we turned 'Jul 1, 2004' into '2004-07-01'.
- Separate each player's personal rating from total rating and convert the data type into numeric type.
- Data type: turn strings into factors as needed.
- Impute goal keepers' missing data in 27 position rating attributes.
- Delete missing data. Since the majority data of 1564 observations is lost and joined data is hard to impute, we chose to delete them all.
- Reset the level of Body.Type. We convert typical ones, such as 'Akinfenwa' and 'C. Ronaldo', into the most frequent level 'Lean', 'Normal', or 'Stocky', based on players actual body type.

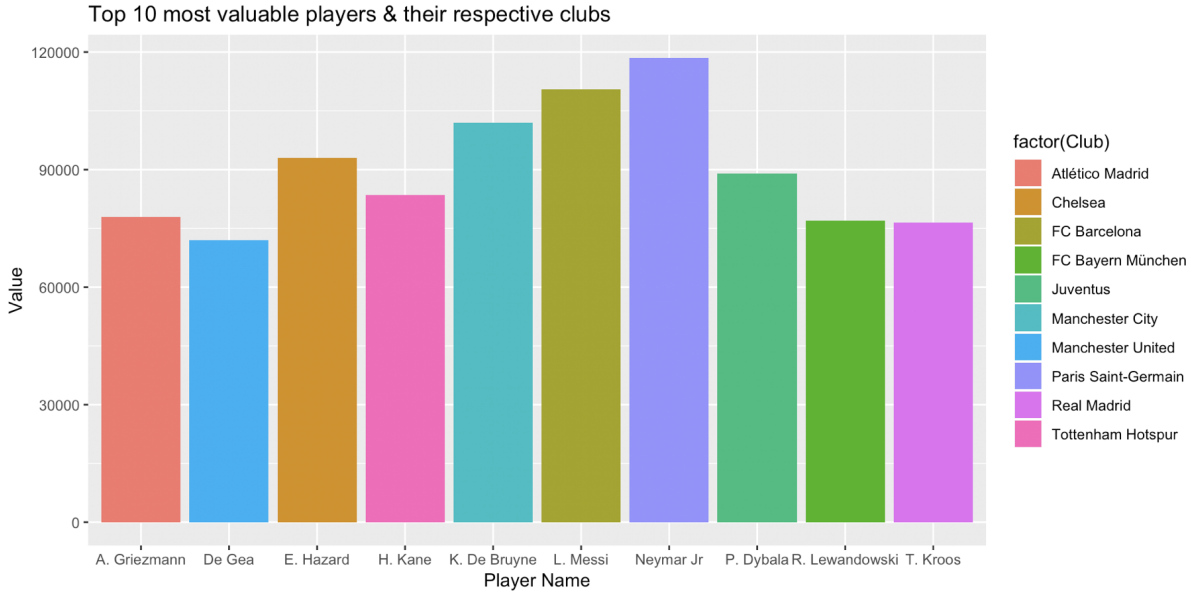
2.2 Data Visualization/EDA

After preparing the dataset for analysis, explorations were made on a macro level like continent and player nationality level analysis as well as on player level. Through the process attempts were made to derive interesting correlations and trends by the use of visualizations.

2.2.1 Exploring top players

By looking at the top players in FIFA, we can get a basic understanding of the dataset, which is a crucial part for further analysis.

1) Plotting top 10 players and their clubs whose value is highest
2) Plotting the superstars in FIFA in each club



2.2.2 Exploring overall ratings

1) Overall Ratings

By plotting histogram chart of overall ratings of players, we spotted that payer ratings are normally distributed in FIFA19, with a mean of 66.2387 and standard deviation of 6.9089.

2) Age vs Overall Rating

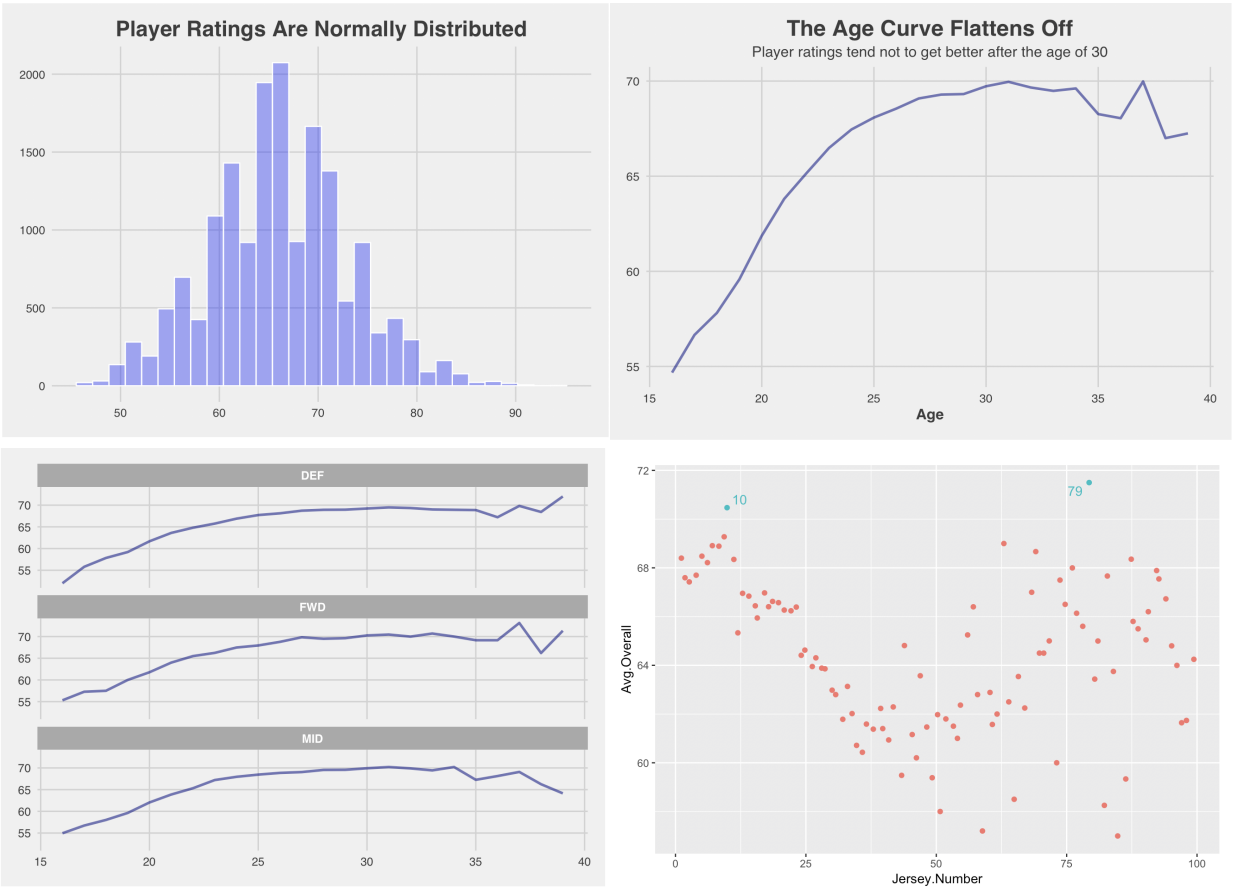
Since age may be an important factor affecting the performance of players, we explore the relationship between players' age and overall rating by drawing a line chart. On average, player ratings stop increasing until around 30 years of age, whereby they level off for a couple of years and then start the inevitable decline at around 34 years.

3) Age vs Overall Rating (Group by positions)

We decided to explore this relationship with the group by variable "positions" because we want to know whether or not age will be an impactor for our position prediction.

The decline for defenders' ratings starts earliest at around 33 years of age, and the decline for both attackers and midfielders starts around 35 years of age.

4) Jersey number and Overall Rating
By plotting players' Jersey number and their overall rating, we noticed that number 10 tend to be the most popular number for awarded players



2.2.3 Exploring player value

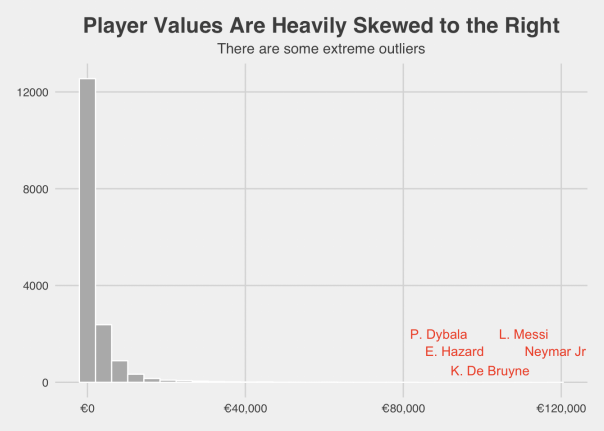
1) Plotting the distribution of player values
By plotting histogram chart for the valuations, we find that player valuations show a heavily positive skew because of the extremely high valuations for the superstars such as Neymar, C. Ronaldo, and Messi. This exploration tells us that we need to scale the data first before we do the clustering.

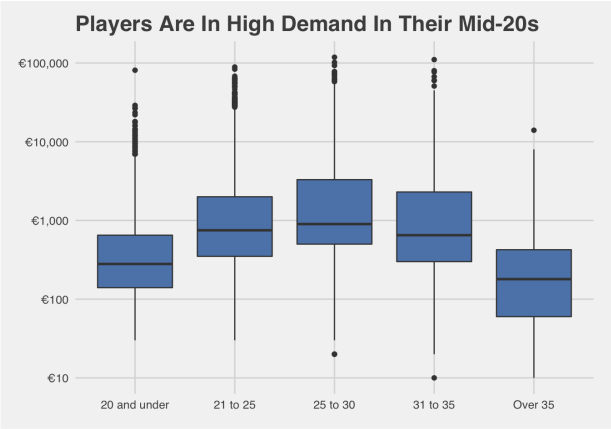
2) Age vs Valuations
Again, we want to explore the relationship between age and valuations. Players' valuation tends to increase up to their early 30s, begin decline around 31, and rapidly decline after 35 years old. The finding matches with the finding for the overall ratings.

3) Positions vs valuations

Here we attempt to show the distribution of player value at different positions.

As we expected, Forwards and Midfielders are going to cost you more than Defenders and Goalkeepers.

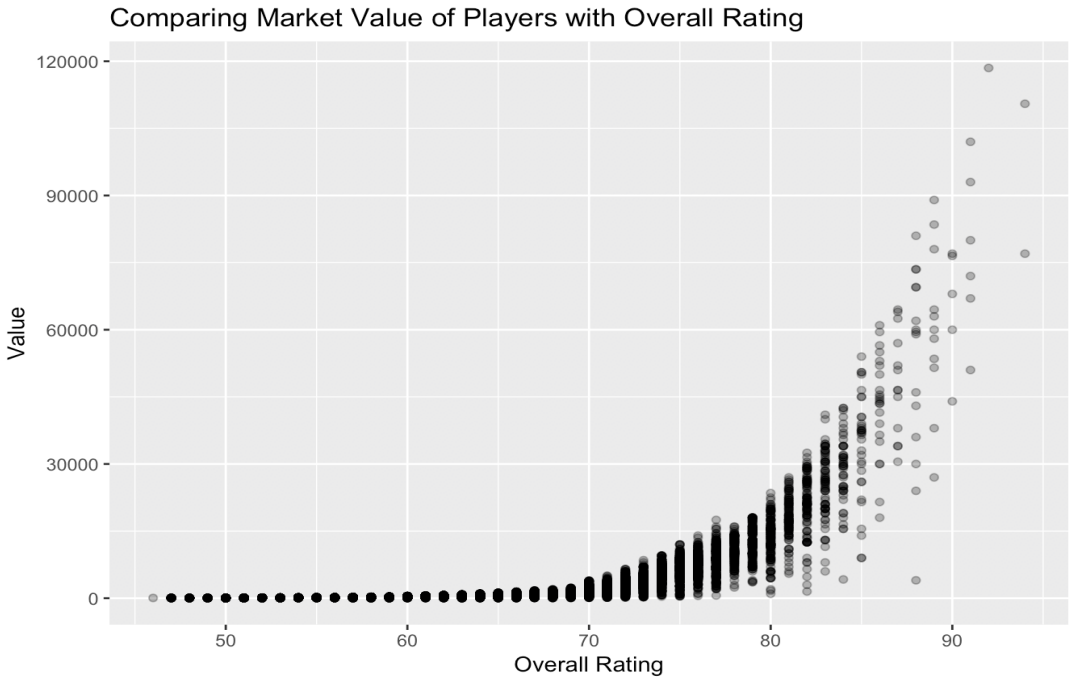




2.2.4 Exploring the relationship between ratings and value

1) Overall Rating and Valuations

We can observe that the value of players increases with the overall rating. However, there are some outliers, whose market values are huge(Messi, Neymar, etc) compared to Overall Rating.



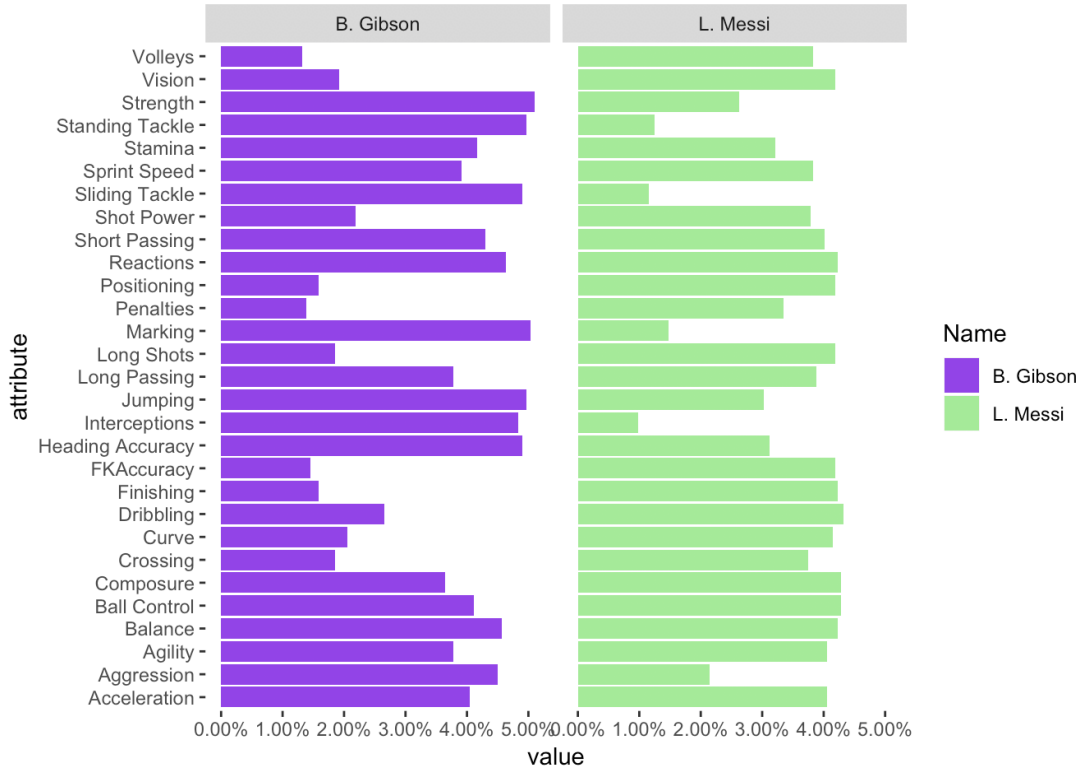
2.2.5 Exploring players attributes

FIFA 2019, Exploring Players Statistics and Build Your Dream Team
Group Member: Suzy Gao, Yufan Luo, Zhijun Liu, Yu Yang

Each player will have particular attributes they excel in compared to other attributes and its this that I want to identify. If you have a particular player who might have lots of paces and a hard shot and you want to find players with similar attributes this will identify those similar players. Therefore, it is very important for us to identify key similarities between different players so that we are able to find potential substitutions for players.

1) A comparison of the percentage of different attributes between two players(Messi & Gibson)

We have chosen Lionel Messi and Ben Gibson, two totally different players to compare their percentages over all the attributes. As you can see, it clearly shows the differences in the players. Ben Gibson has strengths in Jumping, Interceptions and heading accuracy. Whereas Messi clearly stronger at finishing. Therefore, we realized individual strengths are key factors in determining a player's position, and we need to pay more attention to these features when we are building models.



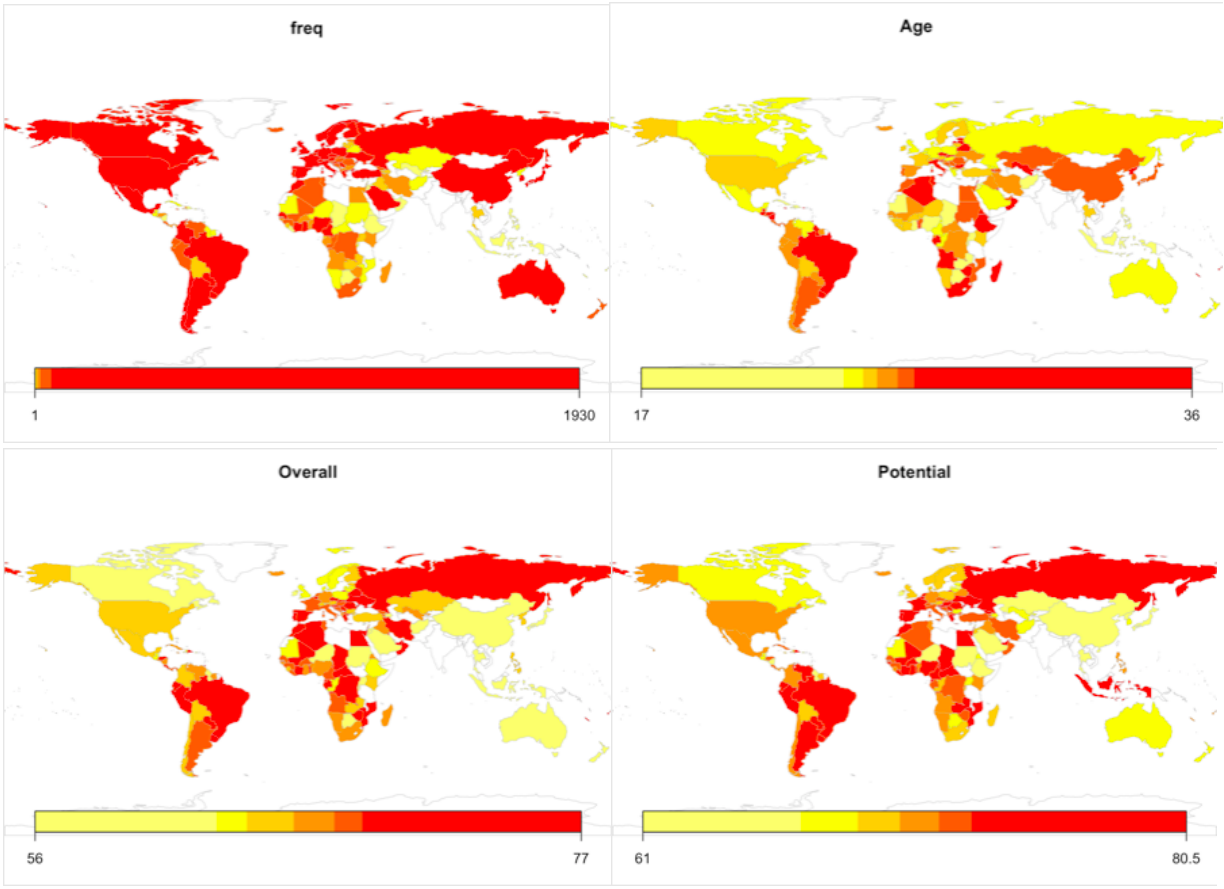
2) Positions and Attributes

By drawing a heat map of players with an overall rating of 75 or higher other than goalkeepers, we found that Right and Left Wingers has great acceleration and aggression, center backs are good at strength, and Left and Right Midfielders are agile.

After doing this exploration, we start to think about building a predictive model to help team manager arrange the most suitable position for each player.

2.2.6 Spatial analysis

Spatial plots can conveniently show the difference among players from different countries. Here we drew four map plots to show players' quantity, average age, average overall rating and average potential rating in each country. We can know that except Africa and West Asia, all continents have a large number of football players. The average age of players in Asia and South America is higher than that of players in other continents. And players in Europe and South America tend to have higher overall and potential ratings.



2.3 Predictive model for position:

For the predictive model, we decided to do clustering first, and build predictive model under each cluster because we noticed that each position requires particular skill sets. By classifying players into relatively homogeneous groups will make our predictions more accurate.

2.3.1 Data preparation

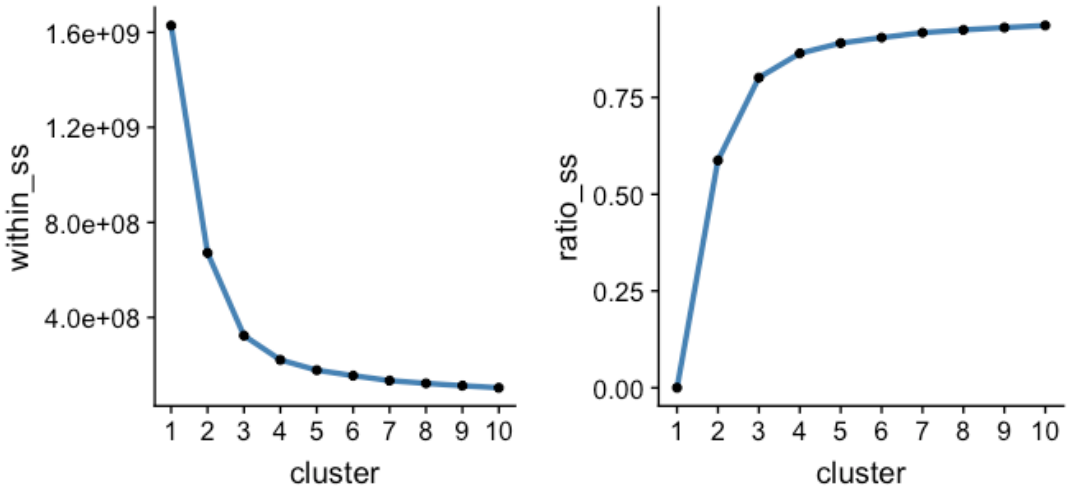
We firstly created a new variable "PositionGroup" to take the place of the original "Position", because FIFA dataset gave detailed classifications of positions. After grouping 27 positions into

4 positions (gk: goal keeper, defs: defenders, mids: midfielders, fwds: forwards), we increased the accuracy of our predictions.

2.3.2 Clustering

In order to increase the final accuracy of predictive models, we used K-Means method to group similar players.

- Because our goal is to predict players' positions, and we wanted to group players based on their skills and physical abilities. Consequently, when doing the clustering, we excluded variables such as, "ID", "Club", "Nationality", "Name", "Value", "Wage", "International.Reputation", "Real.Face", "Jersey.Number", "Joined", "Contract.Valid.Until", "Release.Clause", and "Position".
- For some categorical variables "Preferred.Foot", "Body.Type" and "defend_Work.Rate", we used number to represent their different level.
- We ran K-Means function 10 times based on different number of centers, varying from 1 to 10. The the maximum number of iterations allowed in each function is 10000.
- The following pictures are within sum of squares plot and the ratio plot, and both indicate that the number of clusters should be 2 or 3.



- As a result, we generated two new datasets. The first one has two clusters, and the cluster sizes are 12689 and 3954. The other dataset has three clusters, and the cluster sizes are 7421, 1983 and 7239.

2.3.3 Predictive models:

2.3.3.1 Random Forest

- **Variables**: In the fifa data set, we had position,skills rating and physical strength data about players and we built a random forest model using variables that describe players' skill and physical strength such as 'sliding tackle','finishing' and 'LongPass',etc. We have 27 different specific positions data. Since players in

game and reality usually play more than one role, we group these 27 positions into for Groups: FWD, DEF,GK and MID. FED stands for forwards and GK stands for Goalkeeper. DEF stands for defenders and MID stands for midfielder. Random forest is suitable for this prediction because it can take numeric variables and do classification .

- **Parameters**: We set number of trees as 1000, and the number of variables to possibly split at in each node as 9.
- **Results**: We applied Random Forest model to the entire dataset and each cluster to figure out if the clustering can optimize the prediction. As a result, we found that 2-center-clustering can improve the performance of predictive models.

	Cluster 1	Cluster 2	Average	Entire dataset
Size	12689*0.2	3954 *0.2	-	16643*0.2
Accuracy	0.8611333	0.97875	0.88907637	0.8848855

(Average is the weighted average of all clusters' accuracy, 0.2 is the share of observations in test set)

	Cluster 1	Cluster 2	Cluster 3	Average	Entire Dataset
Size	7421*0.2	1983*0.2	7239*0.2	-	16643*0.2
Accuracy	0.8474462	1	0.8907216	0.88444583	0.8848855

(Average is the weighted average of all clusters' accuracy, 0.2 is the share of observations in test set)

2.3.3.2 Neural Network

The reason for us to choose Neural Network model is that this model is suitable to handle large number of data sets. And it can help to detect the nonlinear relations between variables.

- **Variables**: For Neural Network model, we used all variables except "ID", "Club", "Nationality", "Name", "Real.Face", "Jersey.Number", "Joined", and "Contract.Valid.Until". Because some of the variables are meaningless for us to predict positions, and we also removed variables like "Club" and "Nationality" have hundreds of levels, which could make the prediction difficult.
- **Parameters**: We set input drop ratio is 0.1. Our Neural Network model has there hidden layers, their sizes are (10, 10, 5), and drop ratio for each hidden layer is 0.05, helping to prevent overfitting. The activation function we select is Rectifier with dropout and we oversample the minority to balance the class distribution.

- **Results**: We applied Neural Network model to the entire dataset and each cluster to figure out if the clustering can optimize the prediction. As a result, we found that 2-center-clustering can improve the performance of predictive models.

	Cluster 1	Cluster 2	Average	Entire dataset
Size	12689*0.2	3954*0.2	-	16643
Accuracy	0.8627294	0.98	0.89059024	0.8854962

(Average is the weighted average of all clusters' accuracy,0.2 is the share of observations in test set)

	Cluster 1	Cluster 2	Cluster 3	Average	Entire Dataset
Size	7421*0.2	1983*0.2	7239*0.2	-	16643
Accuracy	0.8487903	1	0.8879725	0.88384941	0.8854962

(Average is the weighted average of all clusters' accuracy,0.2 is the share of observations in test set)

2.3.3.3 Selected Models

We chose the model with higher accuracy for each cluster and compared both method of two clusters with method of three clusters. Finally, we found that combination of Neural Network and K-Means clustering with 2 centers has a better performance on predicting players' positions.

	Cluster 1	Cluster 2	Average
Model	Neural Network	Neural Network	-
Accuracy	0.8627294	0.98	0.89059024

(Average is the weighted average of all clusters' accuracy,0.2 is the share of observations in test set)

	Cluster 1	Cluster 2	Cluster 3	Average
Model	Neural Network	Neural Network/ Random Forest	Random Forest	-
Accuracy	0.8487903	1	0.8907216	0.88504515

(Average is the weighted average of all clusters' accuracy,0.2 is the share of observations in test set)

2.3.4 Skill sets insight & Replacement model:

For the replacement model, we realized that it is difficult to find players who are similar from over 18000 players using over 100 variables. We used K-means clustering because K-means clustering aims to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean. We also use random forest model because it can conduct a importance list, which can help us determine which attributes are important to different position. In the end , we create a function which can recommend players for team manager based on players their need and their budget. The steps used are outlined below:

1) Filter Unknown players (players without a position listed);
2) Since we want to do the clusters based on players' skill sets, we select numeric variables, mainly the attribute variables, and exclude variables such as player value, wages and overall ratings to avoid distractions; We dummy the Position Group variable so that we can unlock insight for each position.
3) Use random forest to unlock important purity for each attribute.
4) Create three boxplot to show the skill set different within different position
5) Build random forest model to predict specify position
6) Build recommendation function

Results:

1. Model prediction accuracy: Our model can predict player for each position accurately, which means the important plots we provide in next step are very reliable.

	Cluster 1	Cluster 2	Average	Entire dataset
Size	12689*0.2	3954 *0.2	-	16643*0.2
Accuracy	0.8611333	0.97875	0.88907637	0.8848855

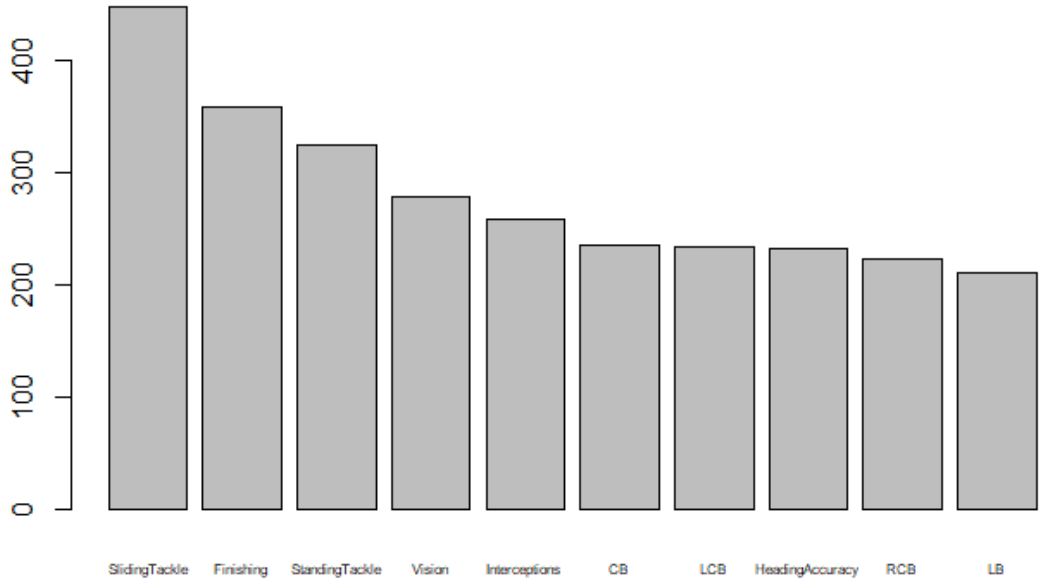
(Average is the weighted average of all clusters' accuracy,0.2 is the share of observations in test set)

	Cluster 1	Cluster 2	Cluster 3	Average	Entire Dataset
Size	7421*0.2	1983*0.2	7239*0.2	-	16643*0.2
Accuracy	0.8474462	1	0.8907216	0.88444583	0.8848855

(Average is the weighted average of all clusters' accuracy,0.2 is the share of observations in test set)

FIFA 2019, Exploring Players Statistics and Build Your Dream Team
Group Member: Suzy Gao, Yufan Luo, Zhijun Liu, Yu Yang

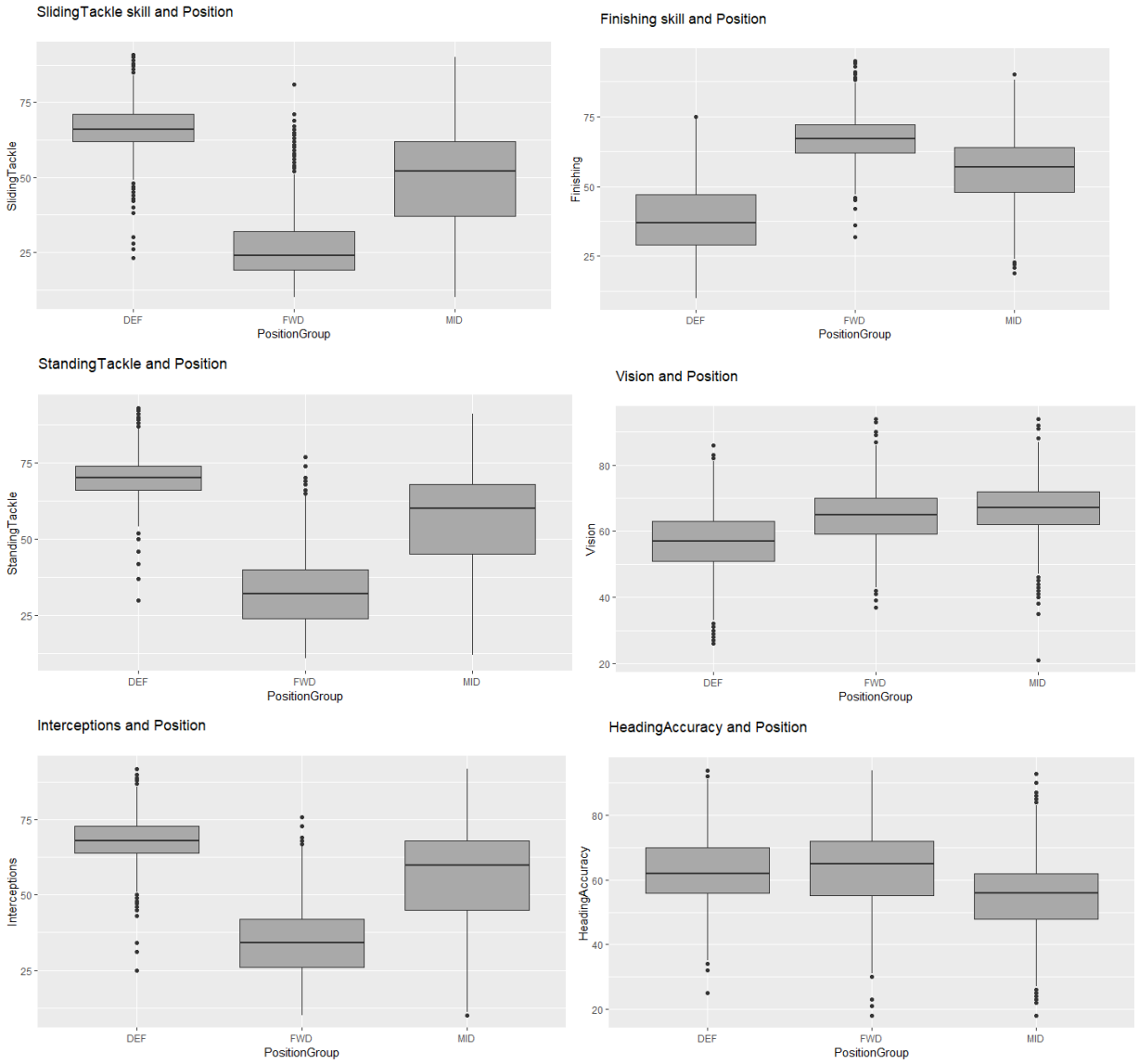
2. Important Plot: We draw a important plot for predicting positions



It seems that players in different positions have very differently skill sets in Sliding Tackle, Finishing, Standing Tackle, Vision, Interceptions and heading accuracy.

3. Difference between different positions: Based on the result of our previous work, Sliding Tackle, Finishing, Standing Tackle, Vision, Interceptions and heading accuracy are important when predicting Position. Therefore, FWD,DEF and MID must behave very differently on these skill sets. We visualize players data on these variables.

FIFA 2019, Exploring Players Statistics and Build Your Dream Team
Group Member: Suzy Gao, Yufan Luo, Zhijun Liu, Yu Yang



It appears that a DEF should have strong Sliding Tackle, Standing Tackle, heading accuracy and Interceptions. FWD players master extremely strong Finishing skill. Their main goal in game is to score. MID players are very versatile. That's partly explain that why MID prediction has the lowest prediction accuracy.

4. Replacement Recommendation: When a soccer manager tries to manage a team, budget usually would be their biggest enemies. In order to solve this problem, we calculate the similarity of each player based on the core_skill list and build a function called findsimilarplayer() to do the recommendation.

Here is the example. Let's say we need a player who plays like Messi(ID:158023), but we only have 57000K dollars,which is far lower than Messi's value- 110500K dollars. Who should we buy for our team?

We just type in Messi's Id and our budget in it like this:

```
findsimilarplayer(158023,57000)
```

Then it can return the top 10 players who are similar to Messi.

	ID <int>	name <fctr>	value <int>	similarity_diatance <dbl>	Potential <int>
1	202556	M. Depay	42000	23.64318	89
2	41236	Z. Ibrahimovi	14000	25.21904	85
3	176769	Jonas	16500	25.25866	84
4	159261	F. Quagliarella	8000	26.87006	81
5	113422	David Villa	8000	26.94439	82
6	172114	D. Valeri	11500	27.33130	80

Based on potential and similarity distance, it seems like M.Depay is our guy!

3. Conclusions and Recommendation

This uninterpreted data can be converted into information by analysing it. By visualizing some of the key features and looking at their relationships, we were able to get a comprehensive understanding of the dataset, which is beneficial for further analysis and modeling process. Insights and correlations between player value, wage, age, special attributes, and performance can be derived from the dataset.For example, we found that a player's attributes could differ a lot from the other, and players in different positions could have entirely different strengths most of the cases, still, they could have some similarities as well. That is to say, player attributes are definitely the key factor that we want to include in the following steps of the analysis.

Position predictive model serves as a reference for managers and facilitates the flexibility against the other teams. Managers use our model can react accordingly by positioning players on the most appropriate positions in case that the against teams have different positioning models. Players' condition may vary from time to time, whereby adopting our most updated model will provide more accurate simulation and greater chance to win. Clustering can group similar players together so that we can make a better prediction in each cluster. For example, grouping all players into two clusters and using Neural Network model to do the prediction in each cluster has the highest accuracy (0.89059024), while the accuracy of using Neural Network model to predict without clustering is 0.8854962.

By Utilizing the random forest importance plot, we explore some insight. we figure out that plays in different position behave very differently in Sliding Tackle, Finishing, Standing Tackle, Vision, Interceptions and heading accuracy. We draw box plot about it to get more insight and find out that DEF has stronger Sliding Tackle, Standing Tackle, heading accuracy and Interceptions skills. FWD has extremely high finishing skills. Midfielder are very versatile and they need to master lots of different skills.

Replacement model enables managers to find a list of replacements for certain players in case that the manger has limited budget for hiring those players. Or the manager can find bench players when some players are absent.

We hope our report will depict a general picture for the overall situation of soccer players. Readers can use our models to gain some insights on how a player is positioned and how to find substitutes for a

certain player. The reports can be used as a reference when game players choose their teams, when EA designs the following FIFA games, and when real-world clubs make their teams.