FIFA Football Player Analysis
Project Report

Claudia Vargas (cv2400)
Kexin Mao (km3404)
Tianying Gong (tg2673)
Yvonne Wang (qw2259)

Table of Contents

Chapter 1 Problem Statement

1.1 Choice for the Scenario

1.2 Problem Statement

Chapter 2 Proposal

2.1 Database Selection

2.2 Work Plan

Chapter 3 Team Structure and Timeline

3.1 Team structure and responsibility

3.2 Timeline

Chapter 4 Database Schema

4.1 Normalization Plan

4.2 ER Diagram

ER Diagram: **https://drive.google.com/open?id=1jPyy_CaHnOr0iqgnv3RsC-FaKHPbovFE**

Chapter 5 Tables Creation

Chapter 6 ETL Process

6.1 Extract

6.2 Transform

6.3 Load

ETL Link, Jupyter Notebook:
https://drive.google.com/open?id=1cziCtlLsbNrjiNJ65PQOSM2szo-iUGDK

Chapter 7 Analytics Applications

7.1 Descriptive Analysis on Value,wages,ratings

7.2 Influence of Age

7.3 Play on Loan

7.4 Wage and Performance

7.5 Case Study

Chapter 8 Metabase Dashboard

http://su19server.apan5310.com:3102/public/dashboard/736186bb-8475-467a-bd5c-7e2732c7b3df

Chapter 9 Conclusion

9.1 Goals achieved

9.2 Benefit of RDMS

9.3 Benefit of ETL

9.4 Benefit of Analysis

Chapter 1 Problem Statement

1.1 Choice for the scenario

There are 2.55 quintillion bytes of data generated every day in the world, however most companies and organization are not translating these data into actionable insights. Our team aims to help with those companies struggling to manage their data by designing an effective database. Using database technology to gather, store and process information, those companies and organizations will be able to manage data efficiently and analyze data in a variety of ways.

Currently, our clients are football clubs in FIFA. FIFA is a non-profit organization, which describes itself as the highest international governing body of football. It has played a pioneering role amongst international sporting federations in the area of governance, issuing regulations and reports in close cooperation with legal and accounting specialists.

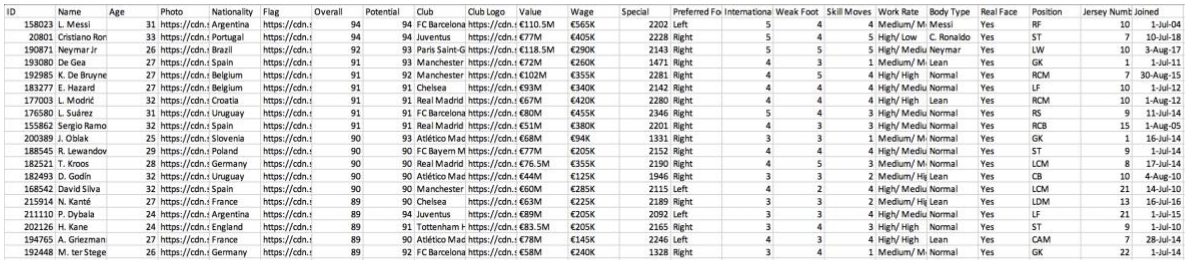
1.2 Problem Statement

The ultimate need from the club is to improve itself. To be more specific, the needs are from two levels of people: team managers and C-suites. The team managers focus on more detailed questions concerning the football player and football team. They would like to better understand the current situation in the football fields and the way to improve the team formation. While, C-suites cares more about the future development of the whole club. They would like to have a better performing team with less cost.

Chapter 2 Proposal

2.1 Database selection

The database we select is a FIFA 2019 database from Kaggle.(Full access: https://www.kaggle.com/karangadiya/fifa19). This FIFA dataset contains 89 variables and 18207 rows. Each row represents a single player. The variables include physical information(weight, height) of each player, club information, market value, and their performance rating for different skills(crossing, long shots) and position. This dataset can help us better understand players’ strengths and weaknesses. Below is a screenshot of part of database.



ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	Club Logo	Value	Wage	Special	Preferred Fo	Internationa	Weak Foot	Skill Moves	Work Rate	Body Type	Real Face	Position	Jersey Numb	Joined
158023	L. Messi	31	https://cdn.s	Argentina	https://cdn.s	94	94	FC Barcelona	https://cdn.s	€110.5M	€565K	2202	Left	5	4	4	Medium/ M	Messi	Yes	RF	10	1-Jul-04
20801	Cristiano Ron	33	https://cdn.s	Portugal	https://cdn.s	94	94	Juventus	https://cdn.s	€77M	€405K	2228	Right	5	4	5	High/ Low	C. Ronaldo	Yes	ST	7	10-Jul-18
190871	Neymar Jr	26	https://cdn.s	Brazil	https://cdn.s	92	93	Paris Saint-G	https://cdn.s	€118.5M	€290K	2143	Right	5	5	5	High/ Mediu	Neymar	Yes	LW	10	3-Aug-17
193080	De Gea	27	https://cdn.s	Spain	https://cdn.s	91	93	Manchester	https://cdn.s	€72M	€260K	1471	Right	4	3	1	Medium/ M	Lean	Yes	GK	1	1-Jul-11
192985	K. De Bruyne	27	https://cdn.s	Belgium	https://cdn.s	91	92	Manchester	https://cdn.s	€102M	€355K	2281	Right	4	5	4	High/ High	Normal	Yes	RCM	7	30-Aug-15
183277	E. Hazard	27	https://cdn.s	Belgium	https://cdn.s	91	91	Chelsea	https://cdn.s	€93M	€340K	2142	Right	4	4	4	High/ Mediu	Normal	Yes	LF	10	1-Jul-12
177003	L. Modrić	32	https://cdn.s	Croatia	https://cdn.s	91	91	Real Madrid	https://cdn.s	€67M	€420K	2280	Right	4	4	4	High/ High	Lean	Yes	RCM	10	1-Aug-12
176580	L. Suárez	31	https://cdn.s	Uruguay	https://cdn.s	91	91	FC Barcelona	https://cdn.s	€80M	€455K	2346	Right	5	4	3	High/ Mediu	Normal	Yes	RS	9	11-Jul-14
155862	Sergio Ramo	32	https://cdn.s	Spain	https://cdn.s	91	91	Real Madrid	https://cdn.s	€51M	€380K	2201	Right	4	3	3	High/ Mediu	Normal	Yes	RCB	15	1-Aug-05
200389	J. Oblak	25	https://cdn.s	Slovenia	https://cdn.s	90	93	Atlético Mad	https://cdn.s	€68M	€94K	1331	Right	3	3	1	Medium/ M	Normal	Yes	GK	1	16-Jul-14
188545	R. Lewandov	29	https://cdn.s	Poland	https://cdn.s	90	90	FC Bayern M	https://cdn.s	€77M	€205K	2152	Right	4	4	4	High/ Mediu	Normal	Yes	ST	9	1-Jul-14
182521	T. Kroos	28	https://cdn.s	Germany	https://cdn.s	90	90	Real Madrid	https://cdn.s	€76.5M	€355K	2190	Right	4	5	3	Medium/ M	Normal	Yes	LCM	8	17-Jul-14
182493	D. Godín	32	https://cdn.s	Uruguay	https://cdn.s	90	90	Atlético Mad	https://cdn.s	€44M	€125K	1946	Right	3	3	2	Medium/ Hi	Lean	Yes	CB	10	4-Aug-10
168542	David Silva	32	https://cdn.s	Spain	https://cdn.s	90	90	Manchester	https://cdn.s	€60M	€285K	2115	Left	4	2	4	High/ Mediu	Normal	Yes	LCM	21	14-Jul-10
215914	N. Kanté	27	https://cdn.s	France	https://cdn.s	89	90	Chelsea	https://cdn.s	€63M	€225K	2189	Right	3	3	2	Medium/ Hi	Lean	Yes	LDM	13	16-Jul-16
211110	P. Dybala	24	https://cdn.s	Argentina	https://cdn.s	89	94	Juventus	https://cdn.s	€89M	€205K	2092	Left	3	3	4	High/ Mediu	Normal	Yes	LF	21	1-Jul-15
202126	H. Kane	24	https://cdn.s	England	https://cdn.s	89	91	Tottenham H	https://cdn.s	€83.5M	€205K	2165	Right	3	4	3	High/ High	Normal	Yes	ST	9	1-Jul-10
194765	A. Griezman	27	https://cdn.s	France	https://cdn.s	89	90	Atlético Mad	https://cdn.s	€78M	€145K	2246	Left	4	3	4	High/ High	Lean	Yes	CAM	7	28-Jul-14
192448	M. ter Stege	26	https://cdn.s	Germany	https://cdn.s	89	92	FC Barcelona	https://cdn.s	€58M	€240K	1328	Right	3	4	1	Medium/ M	Normal	Yes	GK	22	1-Jul-14

The main reason of choosing this database is that it is authorized data from https://sofifa.com.It has completed data for evaluating football players from different clubs, from their physical condition(such as height, weight) to their ability for different moves(such as pass and shots). The most beautiful part of this dataset is that it quantifies the ability of each player, so that we can compare them across different clubs. Also, the data is large enough to conduct analysis and generate impactful insights.

Meanwhile, we have done some research on other football data organizations. In reality, these organizations, such as OPTA, rate football players through goal, assist, interception, tackle and pass accuracy. However, it is hard to apply these data into cross-league competitions, since different leagues have different styles (for example, Spanish put more emphasis on attack; while Italian on defense.) FIFA Football is a football game with over 20 years of history, praised by the whole football industry and football lovers. It quantifies each football player using its own scoring system. Thus, we can leverage on this database to compare with different players and make competition strategy or optimize teams' composition.

2.2 Work plan

The first step would be to clean the data and doing normalization, so that it is easier for us to manage the data and understand attributes. Then, we will analyze the value for key players by further exploring the data in python. The last step is to provide insights and actionable recommendations for our clients.

Generally, this research would help club managers make better decisions on building up their teams from player performance and cost-efficiency perspectives. Our research would perform general and detailed analysis on player's performance features, which would assist the manager in learning the market, targeting the ideal players, and evaluating the potential impact of swapping and/or adding a player to a team. Moreover, this research could also guide managers to make the most cost-efficient decisions in the player transfer market. Instead of purely chasing players who have the highest price, analysis and valuation on players performance would help the manager discover undervalued players who might be more suitable for the team.

Chapter 3 Team Structure and Timeline

3.1 Team structure and responsibility

Each team member is voluntarily in charge of one or some parts of the project, and individual responsibilities are listed below. The rest of the project is evenly distributed to each member.

Tianying Gong: normalize data into at least 15 tables at 3NF

Claudia Vargas Romero: develop a relational schema using Python and load data (ETL)Yvonne Wang: develop ERD and showcase 8+ analytical procedures

Kexin Mao: showcase 2+ analytical procedures and build interactive metabase dashboard

All team: presentation (10 slides and 8 min presentation), report (5-7 pages of text)

The person who is in charge of that part should 1) draft things for discussion 2) revise the answer according to group discussion and other team members' suggestions. 3) Make sure the final work is delivered on time. Rest of the team should perform quality check and provide feedback.

3.2 Timeline

Content	**Person in charge**	**Deadline**
Explanation of variables	Tianying Gong	July 24th.
Normalization	Tianying Gong	July 25th.
ER Diagram	Yvonne Wang	July 28th.
Database Schema Design	Claudia Vargas Romero	July 29th.
ETL	Claudia Vargas Romero	July 31st.
Analytical procedures	Yvonne Wang (8+) Kexin Mao (2+)	August 8th.
Interactive metabase dashboard	Kexin Mao	August 10th.
Report and presentation preparation	Tianying Gong Claudia Vargas Romero Yvonne Wang Kexin Mao	August 10th.

Chapter 4 Database Schema

4.1 Normalization

For the 1st normalization forms, there are two requirements must be met: 1) Domains of all table attributes must be atomic 2) There cannot be repeating attributes. The first requirement is not met in the original database because the Work Rate variable is not atomic. It contains two values in one column linked with a slash. The former one is work rate of attacking and the latter one is work rate of defense. Thus, this column is split into two different columns: attacking_workrate, defense_workrate. The second requirement is also not met because some position variables are repeating attributes with the same value. Values in certain adjacent position columns are the same. For example, all the three attributes LDM, CDM, RDM describe the performance of a player in the position of Center defensive midfielder with exactly the same score. Thus, we only keep one column CDM and delete the other two.

For 2nd normalization forms, one extra requirement must be met on the basis of 1st normalization forms: every non-key attribute must be fully dependent on the key. In order to meet the requirement for 2nd normalization forms, we add the primary key, which contains only one attribute, to all the tables.

For 3rd normalization forms, every non-key attribute must be non-transitively dependent on the key. This requirement is not met in the original database because attribute flag depends on nationality rather than on player_id directly, thus a new table flag should be separated. The same situation happened on attribute club_logo and club, thus a new table club_logo is separated.

According to the requirements listed above, the original database is separated into 22 3rdnormalization forms.

4.2 ER diagram

The ER diagram can be accessed here:
https://drive.google.com/open?id=1jPyy_CaHnOr0iqgnv3RsC-FaKHPbovFE

Please, note that data types and Null options were not included to help with readability of the model. If that information wants to be accessed, you can find it in the table’s creation section included in the ETL file.

Chapter 5 Tables Creation

We wrote some code to create 22 tables corresponding to the normalization plan. We were careful referencing primary keys, foreign keys, data types, and NULL values according to the normalization plan and the data that was going to be loaded. Also, we implemented many check constraints to make sure that the values entered in the dataset correspond to the values expected for each attribute.

Chapter 6 ETL Process

Full ETL process can be accessed through:
https://drive.google.com/open?id=1cziCtlLsbNrjiNJ65PQOSM2szo-iUGDK

6.1 Extract

The extraction process was performed on a FIFA dataset consisting of 18,207 rows. Each row provides information for a different soccer players around the world. The FIFA dataset contains 89 columns; each column provides information regarding a different characteristic for each soccer player.

6.2 Transform

Missing Values:

One of the main tasks, corresponding to the transformation process, was to process the variables (or attributes) that contained missing values. In order to deal appropriately with them, it was decided to input the ‘No provided’ string value if the variable was of type string or date, and input the -99 number value if the variable was of numeric type. By doing so, it was possible to transform and clean the whole dataset.

Separation of Variables:

Variables “Work Rate”, “Height”, “lwb” (left wing back), “rwb” (right wing back), “lb”(left back), “cb” (center back), “rb” (right back), “st” (striker), “lw” (left winger), “cf” (center forward), “rw” (right winger), “cam” (center attacking midfielder), “lm” (left midfielder), “cm” (center midfielder), “rm” (right midfielder), and “cdm” (center defensive midfielder) contained two values in each row which needed to be separated.

Work Rate: This variable contained values in each row such as (‘High/ Low’). The first value corresponded to the players’ attacking work rate. The second value corresponded to the players’ defense work rate. Hence, the variable “Work Rate” needed to be split into two different columns: attacking_workrate and defense_workrate.

Height: This variable contained values in each row such as ('5/9). The first value corresponded to the players' height in feet. The second value corresponded to the players' height in inches. Hence, the variable "Height" needed to be split into two different columns: height_feet and height_inches.

The variables in the dataset that consisted of acronyms such as lwb, rwb, etc., contained values in each row such as ('64+2'). The former value is the actual rating of each player for each variable (i.e. position); the latter value represents the potential of each player corresponding to each variable. In our analysis, we only needed to keep the first value and delete the second. That is because we are measuring each players' actual performance and not their potential.

Change String Type Variables to Numeric Type:

Variables such as "Release Clause", "Value", and "Wage" contained values for each player such as '€226.5M'. Since those values are strings type and we wouldn't be able to conduct a numerical analysis on them, we decided to clean them up by transforming them to their numeric value. In this case, the result would be 226500000.

Change Numeric Type Variables to String Type:

The variable "International Reputation" contained values such as 1, 2, 3, 4, 5, and nan. Those numbers correspond to the labels: Normal (1), Regional (2), National (3), Continental (4), and World Class (5). Hence, we decided to change the numbers for their actual meaning so that the dataset will be easier to read and perform analysis on it would be easier too.

Clean Values:

Variable "Contract Until" had as entries values such as "May 31, 2020" and "2023". Hence since we cannot compare those values, we decided to just keep the year of each entry. Hence, part of the information would be missed, but having only the year (when the contract of each player will finish), would allow us to analyze this variable.

The variable "Weight" had entries such as '159lbs'. Since we wouldn't be able to perform analysis on this string variable, we decided to take rid of the 'lbs' string, convert the variable to numeric, and rename this variable with "Weight in Pounds" (weight_pounds).

6.3 Load

After all the data in the FIFA dataset was transformed, we created tables corresponding to each one of the tables created through the normalization process. Also, we created IDs for each one of the tables that needed an ID. After doing that, we changed the imputed missing values (‘No provided’ for string and date types variables and -99 for numeric type variables) to missing values (i.e. NaN). We verified that the newly created tables corresponded to the original dataset and loaded the data accordingly. We were very careful not to load the values that we imputed corresponding to missing values. After loading the data to pgAdmin4, we mapped the IDs we just created back to the transformed FIFA dataset. The transformed FIFA dataset contains the imputed values created to replace the missing values.

Chapter 7 Analytics Applications

7.1 Value,wages,ratings distributed by clubs

Question: What are the top 15 clubs with highest value/wages/overall ratings, and how did they distribute? Is there a correlation between value,wages with overall ratings of clubs?

Step to analysis:

1.Join player table with rating, value, clubs, & create new variables on average and rankings

2.Extract data in python to dataframe and analysis with pandas, numpy, matplotlib, & seaborn

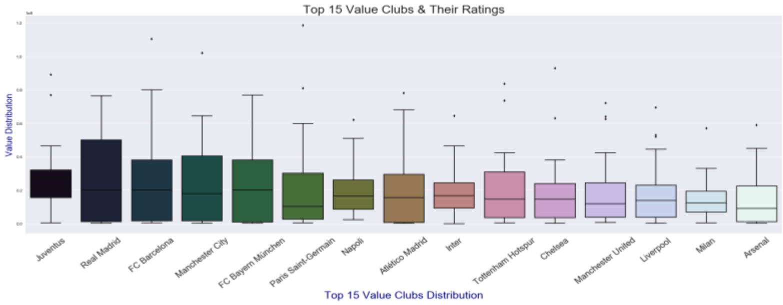
SQL Code:

```
Select * from (SELECT  club_name, player_id, overall, average_overall,
DENSE_RANK()OVER(ORDER BY average_overall DESC)AS overall_rating_rank,
value, average_value, DENSE_RANK()OVER(ORDER BY average_value DESC) AS
value_rank, wage, average_wage, DENSE_RANK()OVER(ORDER BY average_wage
DESC) AS wage_rank, age, average_age, DENSE_RANK()OVER(ORDER BY
average_age ASC) AS age_rank
FROM  (SELECT cb.club_id, r.overall,v.value,v.wage,cb.club_name,p.player_id,p.age,
ROUND(AVG(r.overall)OVER(PARTITION BY cb.club_id),2)AS average_overall,
ROUND(AVG(v.value)OVER(PARTITION BY cb.club_id),0) AS average_value,
ROUND(AVG(v.wage)OVER(PARTITION BY cb.club_id),0) AS average_wage,
ROUND(AVG(p.age)OVER(PARTITION BY cb.club_id),2) AS average_age
FROM player p JOIN club cb ON cb.club_id = p.club_id  JOIN rating r ON r.player_id =
p.player_id Join value v ON v.player_id = p.player_id) sub1) sub2;
```

Findings:

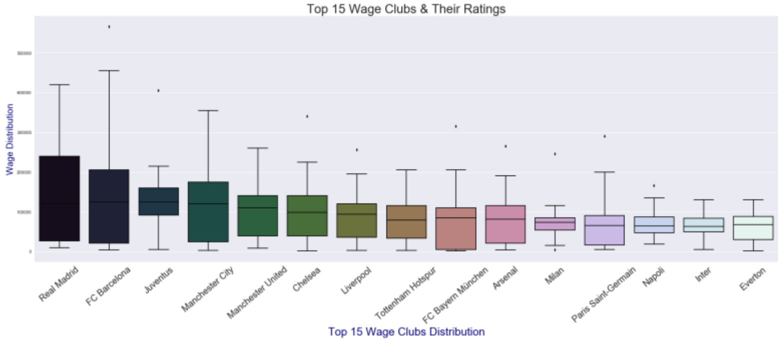
1) Top clubs, distributions, and interesting findings:

1.1 top 15 clubs with highest player value:



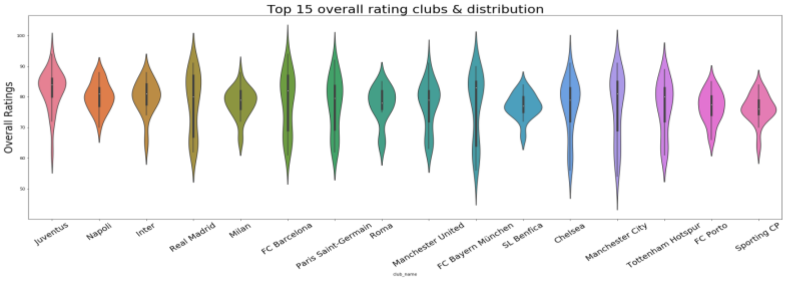
Juventus has the highest average player value and is very central distributed. While generally, the distribution is quite skewed.

1.2 top 15 clubs with highest player wages:



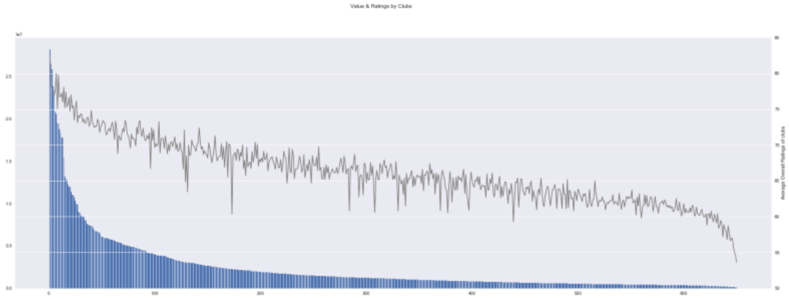
Real Madrid has the highest average wages. Wages are widely distributed for top clubs like Real,Barca,and ManCity, which indicates an unhealthy salary structure of the club.

1.3 top 15 clubs with highest player overall ratings:



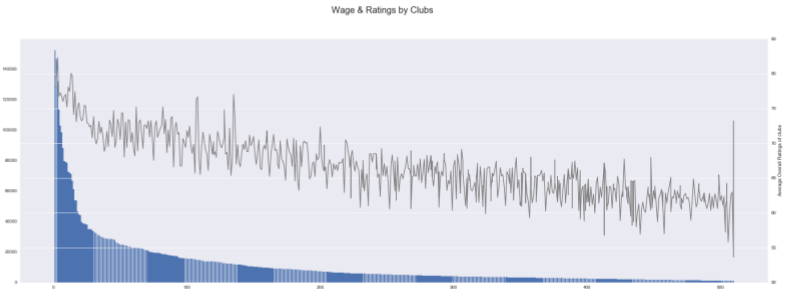
1/3 of the top 15 ratings clubs are traditional Italian clubs with more centralized distributions of player's ratings.

2) Values & Ratings



From Clubs' perspective, the correlation between average value & overall ratings of clubs is 0.78, which indicates it's highly related. *(the barcharts as average value and lines as average ratings.)*

3) Wages & Ratings



From Clubs’ perspective, generally the correlation between average wage & overall ratings of clubs is 0.67, which indicates it’s highly related.
(the barcharts as average wage and lines as average ratings.)

7.2 Value, wages, overall ratings, potential ratings, position performance(by position rating) BY age

Question: How do age influences players’ value, wages, ratings, potentials, and their performances on different positions?

Step to analysis:

1.Join player table with rating, value, clubs, as well as tables for ratings on various positions, eg. forwards, midfielders, defensive & create new variables on average and rankings

2. Breakdown position column to align with player’ rating for that exact position.

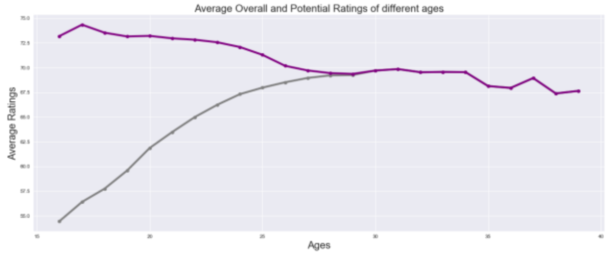
2.Extract data in python to dataframe and analysis with pandas, numpy, matplotlib, & seaborn

SQL code:

```
select distinct *
From (select age, count(age)over(partition by age) as n_age,
round(avg(overall)over(partition by age),2) as avg_overall,
round(avg(potential)over(partition by age),2) as avg_potential,
round(avg(value)over(partition by age),0) as avg_value,
round(avg(wage)over(partition by age),0) as avg_wage,
round(avg(case when position = 'GK' then position_rating end) over (partition by age),
2) as GK_rating, round(avg(case when position = 'LWB' then position_rating  when
position = 'LWB' then position_rating  when position = 'LB' then position_rating when
position = 'CB' then position_rating when position = 'RB' then position_rating end) over
(partition by age), 2) as defensive_rating, round(avg(case when position = 'CAM' then
position_rating when position = 'LM' then position_rating when position = 'CM' then
position_rating when position = 'RM' then position_rating when position = 'CDM' then
position_rating end) over (partition by age), 2) as midfielders_rating, round(avg(case
when position = 'ST' then position_rating when position = 'LW' then position_rating
when position = 'CF' then position_rating when position = 'RW' then position_rating end)
over (partition by age), 2) as forwards_rating
from (select age,position, overall, potential, value, wage,
case when position = 'GK' then overall when position = 'LWB' then lwb when position =
'RWB' then rwb when position = 'LB' then lb when position = 'CB' then cb when position
= 'RB' then rb when position = 'CAM' then cam when position = 'LM' then lm when
position = 'CM' then lb when position = 'RM' then rm when position = 'CDM' then cdm
when position = 'ST' then st when position = 'LW' then lw when position = 'CF' then cf
when position = 'RW' then rw end as position_rating
From (select  p.age,p.position, r.overall,r.potential,v.value,v.wage,
Def.lwb,def.rwb,def.lb,def.cb,def.rb, mid.cam,mid.lm,mid.cm,mid.rm,mid.cdm,
f.st,f.lw,f.cf,f.rw
from player p join rating r on r.player_id = p.player_id join value v on v.player_id =
p.player_id
join defensive def on def.player_id = p.player_id join forwards f on f.player_id  =
p.player_id
join midfielders mid on mid.player_id = p.player_id)SUB1)SUB2)sub3 order by age;
```

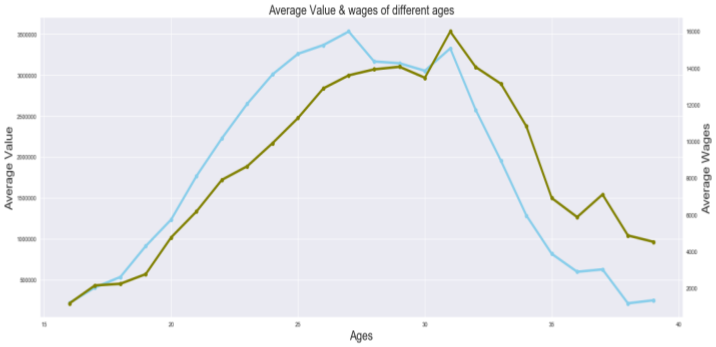
Findings:

1) Age influence on Potential & Overall Ratings



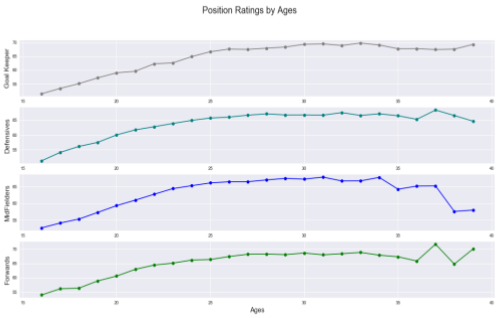
Player’s overall ability*(grey line)* increases with age until 30s while potential ability*(purple line)* decreases with age until 30, and remains stable afterwards.

2) Age influence on Value and Wages



The trend of player’s wage and values by ages are nearly the same.
Generally, wages and value increase before 28-31, and decrease afterwards.
The peak of wages comes 3-4 years later than the peak of value.Age 28 - 31 might be the peak for most player’s career path.

3) Age influence on Performances in different positions



It’s important to choose the best player of the most suitable ages. For players of different positions, their performances generally grow with ages till 30s. Midfielders’ performances*(blue line)* tend to decrease more greatly while Goalkeeper’s performances*(the upper first line)* remain stable.

7.3. Player on loan

Question: What positions are most needed and least needed for clubs that need to borrow players?

SQL CODE:

```
select count(x.position), x.position from
(select player.player_id, player.position,player.age, player.name,club.club_id,
club.club_name,contract.loaned_from, rating.overall, rating.potential from player
join club on player.club_id = club.club_id
join contract on player.contract_id = contract.contract_id
join rating on player.player_id = rating.player_id
where loaned_from is not NUll)x
group by x.position
```

Findings:

1) Most needed: ST(206 clubs), CB(105 clubs), CM(101 clubs) and least needed: LAM(1 club), LWB(2 clubs), RWB(3 clubs)
   We can tell that most clubs that borrow players are looking for assistance in attacking

Question: Normally, we borrow players who play well to help the club. This idea seems obvious but let’s test our assumption.

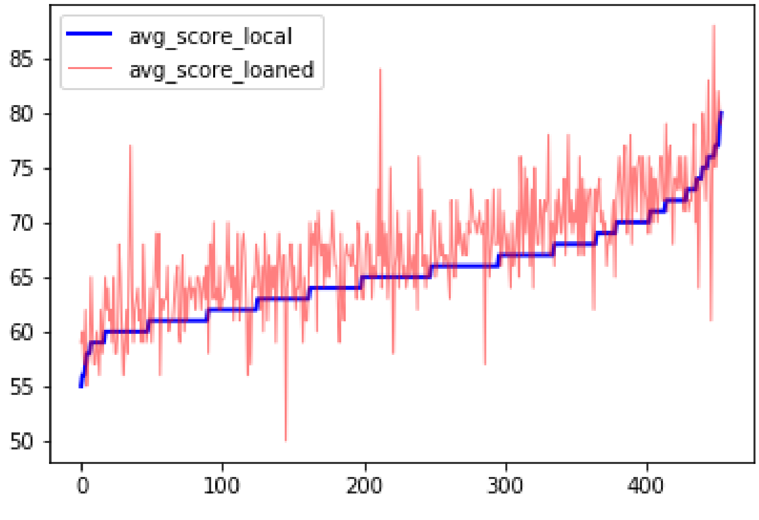
Steps to analyze:

1) To join player, club, contract, rating together

2) To find clubs that have loaned players from other clubs and calculate average score at the club level

SQL CODE:

```
WITH tmp as
(
select club_score_local.*, club_score_loaned.avg_score_loaned from
(select club_id, sum(overall)/count(*) as avg_score_local from
(select * from
(select player.player_id,player.name,club.club_id,
club.club_name,contract.loaned_from, rating.overall, rating.potential from player
join club on player.club_id = club.club_id
join contract on player.contract_id = contract.contract_id
join rating on player.player_id = rating.player_id) x
where x.loaned_from is NUll) local_players
group by 1) club_score_local
left join
(select club_id, sum(overall)/count(*) as avg_score_loaned from
(select * from
(select player.player_id,player.name,club.club_id,
club.club_name,contract.loaned_from, rating.overall, rating.potential from player
join club on player.club_id = club.club_id
join contract on player.contract_id = contract.contract_id
join rating on player.player_id = rating.player_id) x
where x.loaned_from is not NUll) loaned_players
group by 1) club_score_loaned
on club_score_local.club_id = club_score_loaned.club_id
where avg_score_loaned is not NULL )
select avg(avg_score_loaned - avg_score_local) as avg_diff from tmp
```



Findings:

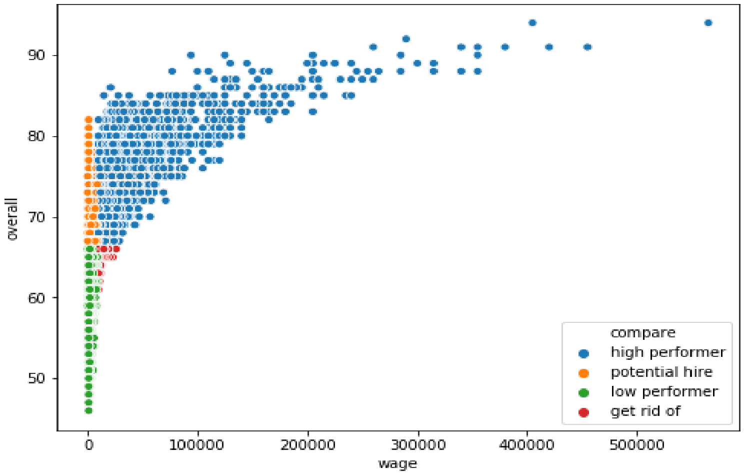
1) 454 clubs have players on loan.
2) On average, the score of players on loan is 2.06 higher than the score of regular players.

7.4. wage and performance — potential hire

Question: Sometimes the wage and performance do not match. As club manager, I would like to hire people with good performance and ideally these people won't require too much money.

SQL CODE:

```
select z.*, case when wage > avg_wage and overall > avg_overall then 'high performer'
        when wage <= avg_wage and overall <= avg_overall then 'low performer'
                when wage > avg_wage and overall <= avg_overall then 'get rid of'
                else 'potential hire' end as compare from
(select x.*, y.* from
(select r.player_id, r.overall, v.wage
from rating as r
join value as v on r.player_id = v.player_id) x
CROSS join
(select sum(overall)/count(*) as avg_overall, sum(wage)/count(*) as avg_wage
from rating as r
join value as v on r.player_id = v.player_id) y)z
```



Findings:

1) The blue dots are the ones who get high wages and have good performance.
2) The green dots are the ones with low wages and low overall score.
3) The yellow dots are the ones who get low wage, however, play well. These are the people we probably want to hire if needed.
4) The red dots are the players who require high wage, but did not do well, and we probably want to avoid these people and at the same time we should be aware if anyone in the club falls into such range. (Here we set the average wage and average overall score as baseline. If above we consider it to be high, and vice versa)

7.5. Case Study - Top clubs

Steps to analyze:

1) To calculate the average score from many aspects (dribbling, defending, goalkeeping, passing, shooting and physical) .
2) To average score for each aspect is calculated from the sum score of many tests to measure such aspect.
3) To join tables together and calculate above average percentage score, so the positive result means above average and negative score means below average.

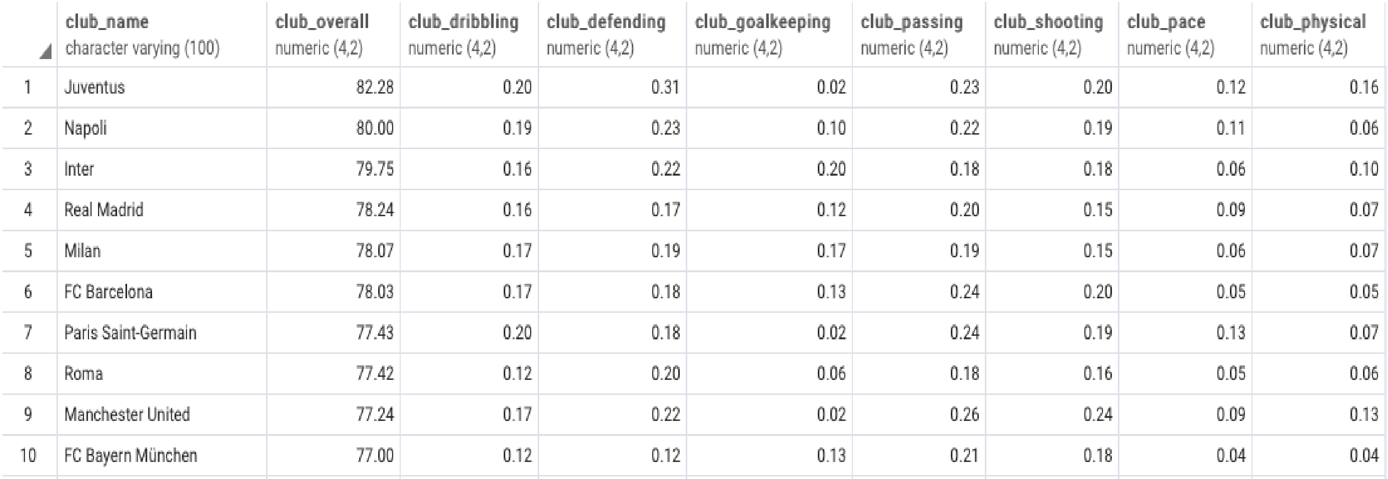
SQL Code:

```
select sum(dribbling + ballcontrol+ agility + reactions + balance +
composure)/count(*)as avg_dribbling_score from dribbling
select sum(heading_accuracy + interceptions + marking + standing_tackle +
sliding_tackle)/count(*) as avg_defending from defending
select sum(gkdiving + gkhandling + gkkicking + gkpositioning + gkreflexes)/count(*) as
avg_goalkeeping_score from goal_keeping
select sum(crossing + shortpassing + curve + fkaccuracy + longpassing +
vision)/count(*) as avg_passing_score from passing
select sum(finishing + volleys + shotpower + longshots + positioning +
penalties)/count(*) as avg_shooting_score from shooting
select sum(acceleration + sprint_speed)/count(*) as avg_pace_score from pace
select sum(jumping + stamina + strength + aggression)/count(*) as avg_physical_score
from physical
```

Result:

```
avg_ dribbling_score = 360
avg_defending_score = 239
avg_goalkeeping_score = 82
avg_passing_score = 303
avg_shooting_score = 288
avg_pace_score = 129
avg_physical_score = 248
```

```
WITH tmp as
(
select x.club_name,x.player_id, x.position, x.overall,
(x.dribbling+x.ballcontrol+x.agility+x.reactions+x.balance+x.composure)/360.0 -1 as
dribbling_score,
(x.heading_accuracy+x.interceptions+x.marking+x.standing_tackle+x.sliding_tackle)/23
9.0 -1 as defending_score,
(x.gkdiving+ x.gkhandling+ x.gkkicking+ x.gkpositioning+ x.gkreflexes)/82.0 -1 as
goalkeeping_score,
(x.crossing+x.shortpassing+x.curve+x.fkaccuracy+x.longpassing+x.vision)/303.0 -1 as
passing_score,
(x.finishing + x.volleys + x.shotpower + x.longshots + x.positioning + x.penalties)/288.0 -
1 as shooting_score,
(x.acceleration+ x.sprint_speed)/129.0 -1 as pace_score,
(x.jumping + x.stamina + x.strength + x.aggression)/ 248.0 -1 as physical_score
from
(select
par.*,dri.dribbling,dri.ballcontrol,dri.agility,dri.reactions,dri.balance,dri.composure,
 def.heading_accuracy, def.interceptions, def.marking, def.standing_tackle,
def.sliding_tackle,
 goal.gkdiving, goal.gkhandling, goal.gkkicking, goal.gkpositioning, goal.gkreflexes,
 pass.crossing, pass.shortpassing, pass.curve, pass.fkaccuracy, pass.longpassing,
pass.vision,
 shoot.finishing, shoot.volleys,shoot.shotpower,shoot.longshots,shoot.positioning,
shoot.penalties,
 pace.acceleration, pace.sprint_speed,
 phy.jumping, phy.stamina, phy.strength, phy.aggression,
 rt.overall,
 c.club_name,
 p.position
from player_ability_rating as par
join dribbling as dri on dri.dribbling_id = par.dribbling_id
join defending as def on def.defending_id = par.defending_id
join goal_keeping as goal on goal.goalkeeping_id = par.goalkeeping_id
join passing as pass on pass.passing_id = par.passing_id
join shooting as shoot on shoot.shooting_id = par.shooting_id
join pace on pace.pace_id = par.pace_id
join physical as phy on phy.physical_id = par.physical_id
join rating as rt on rt.player_id = par.player_id
join player as p on par.player_id = p.player_id
join club as c on c.club_id = p.club_id)x
)
select tmp.club_name, cast(avg(overall)as numeric(4,2)) as club_overall,
cast(avg(dribbling_score)as numeric(4,2)) club_dribbling, cast(avg(defending_score) as
numeric(4,2))as club_defending,
cast(avg(goalkeeping_score)as numeric(4,2)) as club_goalkeeping,
cast(avg(passing_score) as numeric(4,2)) as club_passing,
cast(avg(shooting_score)as numeric(4,2)) as club_shooting, cast(avg(pace_score)as
numeric(4,2))as club_pace,
cast(avg(physical_score) as numeric(4,2)) as club_physical
from tmp
group by tmp.club_name
```



	club_name character varying (100)	club_overall numeric (4,2)	club_dribbling numeric (4,2)	club_defending numeric (4,2)	club_goalkeeping numeric (4,2)	club_passing numeric (4,2)	club_shooting numeric (4,2)	club_pace numeric (4,2)	club_physical numeric (4,2)
1	Juventus	82.28	0.20	0.31	0.02	0.23	0.20	0.12	0.16
2	Napoli	80.00	0.19	0.23	0.10	0.22	0.19	0.11	0.06
3	Inter	79.75	0.16	0.22	0.20	0.18	0.18	0.06	0.10
4	Real Madrid	78.24	0.16	0.17	0.12	0.20	0.15	0.09	0.07
5	Milan	78.07	0.17	0.19	0.17	0.19	0.15	0.06	0.07
6	FC Barcelona	78.03	0.17	0.18	0.13	0.24	0.20	0.05	0.05
7	Paris Saint-Germain	77.43	0.20	0.18	0.02	0.24	0.19	0.13	0.07
8	Roma	77.42	0.12	0.20	0.06	0.18	0.16	0.05	0.06
9	Manchester United	77.24	0.17	0.22	0.02	0.26	0.24	0.09	0.13
10	FC Bayern München	77.00	0.12	0.12	0.13	0.21	0.18	0.04	0.04

```
order by club_overall
```

Findings:

1) All top clubs have very high overall score and are good at all aspects. None of the score is below average.
2) The physical score of the top clubs is not as impressive as other score, meaning that other skills is the key to success
3) Among all the skills, top clubs are especially good at passing, shooting and defending.

Chapter 8 Metabase Dashboard

8.1 Link to Dashboard:

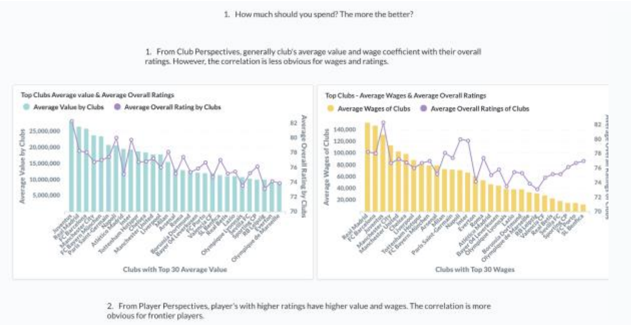
http://su19server.apan5310.com:3102/public/dashboard/736186bb-8475-467a-bd5c-7e2732c7b3df

8.2 Aim of Dashboard:

The aim of the dashboard is to showcase the key analytic findings of the project and provide market-centric and data-driven insights for the clients, current and future soccer team manager on how to build up their team from money, time and team-building perspectives.

8.3 Basic Flow & Explanation:

Part 1. How much should you spend?

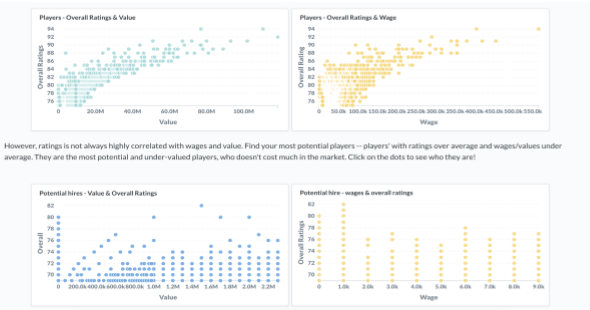


1. How much should you spend? The more the better?

1. From Club Perspectives, generally club's average value and wage coefficient with their overall ratings. However, the correlation is less obvious for wages and ratings.

Top Clubs Average value & Average Overall Ratings
Average Value by Clubs
Average Overall Rating by Clubs
Average Value by Clubs
Average Overall Rating by Clubs
Clubs with Top 30 Average Value
Top Clubs - Average Wages & Average Overall Ratings
Average Wages of Clubs
Average Overall Ratings of Clubs
Average Wages of Clubs
Clubs with Top 30 Wages

2. From Player Perspectives, player's with higher ratings have higher value and wages. The correlation is more obvious for frontier players.

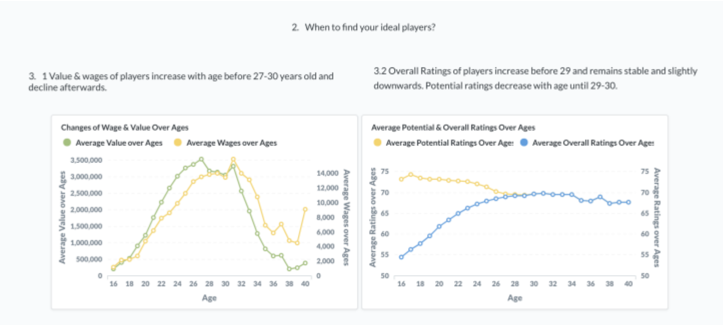


Players - Overall Ratings & Value
Overall Ratings
Value
Players - Overall Ratings & Wage
Overall Rating
Wage
Potential hires - Value & Overall Ratings
Overall
Value
Potential hire - wages & overall ratings
Overall Ratings
Wage

However, ratings is not always highly correlated with wages and value. Find your most potential players -- players' with ratings over average and wages/values under average. They are the most potential and under-valued players, who doesn't cost much in the market. Click on the dots to see who they are!

The first part shows how performances, measured by overall ratings, distributed by wage and value from club average and players individual perspectives. This part help the manager to learn is it really necessary to follow the market trend of spending more money for famous stars. Meanwhile, we also clustered players to a group of highly potential highers, with lower costs but high performances. Fancy In the dashboard, the manager can see a more detailed information about the player by clicking on the spot.

Part 2. When should you find the ideal player?

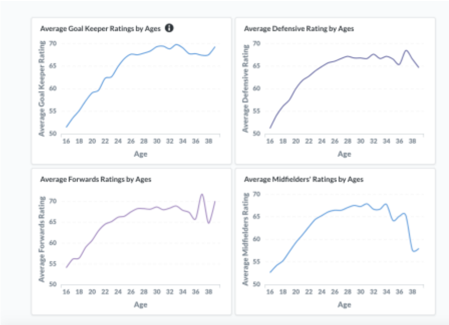


2. When to find your ideal players?

3. 1 Value & wages of players increase with age before 27-30 years old and decline afterwards.

3.2 Overall Ratings of players increase before 29 and remains stable and slightly downwards. Potential ratings decrease with age until 29-30.

Changes of Wage & Value Over Ages
Average Value over Ages
Average Wages over Ages
Average Value over Ages
Average Wages over Ages
Age
Average Potential & Overall Ratings Over Ages
Average Potential Ratings Over Age:
Average Overall Ratings Over Age:
Average Ratings over Ages
Age

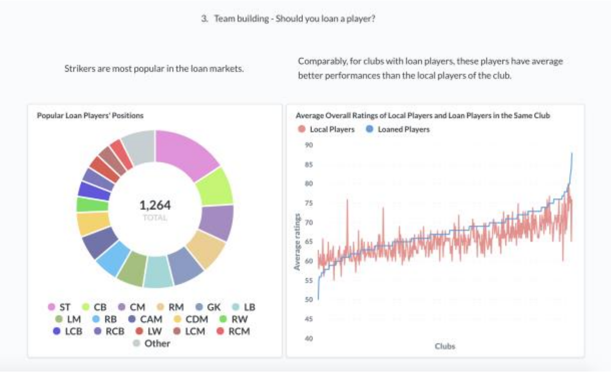


Average Goal Keeper Ratings by Ages
Average Goal Keeper Rating
Age
Average Defensive Rating by Ages
Average Defensive Rating
Age
Average Forwards Ratings by Ages
Average Forwards Rating
Age
Average Midfielders' Ratings by Ages
Average Midfielders Rating
Age

Timing is always important for a players' career path. This is a more interactive showcase of the findings we had in the analysis reports. Managers can refer to the trend of performance and costs changing with time. Moreover, a more detailed illustration on trending pf different position can help them better target the ideal players.

Part 3. Building up the team: Do you need to loan a player?

Besides buying a player, loan player is also a choice for managers. This part of the dashboard shows the most popular positions of players in the loan market. We also compared the average performances of loan players and local players of the same club to see the differences.



3. Team building - Should you loan a player?

Strikers are most popular in the loan markets.

Comparably, for clubs with loan players, these players have average better performances than the local players of the club.

The interesting part about this session is that when moving around the lines, the viewer can see the ratings and club name for their references.

Chapter 9 Conclusion

9.1 Goals achieved

Our goal is to create a database that will help analyze the football player data, in order to provide insights for both team managers and C-suites to make the football team better. For team managers, we would like to give them some basic information concerning the football fields, ways to improve team formation and learning from successful clubs. For C-suites, we would like to provide visualized charts and direct insights for them to run a more successful club and winning more money.

The steps of achieving our goal is through:

1) Build a relational database through normalization.
2) Use ETL to input the original data into the relational database.
3) Link the relational database with Python for data analysis to answer the request from team managers.

4) Link the relational database with Metadata to generate interactive dashboard for C-suites.

9.2 Benefit of RDMS

1) **Well-designed structure**
The database was separated according to different fields. Meanwhile, some variable names are changed to ensure the consistency and better understanding. This makes it easier for both data analysts and clients to use the data.
2) **Improve data integrity**
Data integrity was improved in two ways. Firstly, constraints were set when building the data schema, thus it reduces the possibility of mistakenly input. For example, for preferred foot, we set the constraint to check if the input was in left/right. If the input was not one of these two words, an error will return. Secondly, when original database was separated into different forms, it makes it easier for us to maintain the table, which is to update, insert and delete.
3) **Reduce data redundancy**
RDMS reduces the redundant data. For example, club logo was separated into a different table club and only appear once in that table. While in the original database, this information repeated for every player.

9.3 Benefit of ETL

The ETL process was very helpful because. by conducting it, we gained a very sophisticated knowledge of the FIFA dataset. Knowing very well the dataset had a positive impact on the scope of the project and the kind of questions that we can solve. Also, knowing very well the data enable us to answer very detailed questions to “C” level officers and analysts alike.

9.4 Benefit of Analysis

Analysis is performed both at the player level, and the club level. Both descriptive analysis that shows the distribution of value, wages and rating of top 15 clubs, and in-depth analysis to answer questions like who the players are to hire/avoid are included. In addition, studies like influence of age on performance and case studies of top 10 clubs will help club managers understand patterns in players’ lifecycle and learn from the top clubs.
We have more interactive charts and showcases in the metadata dashboard to help better understand the player market and building up an ideal team.