Report of European Soccer Database

-Dataset

the dataset i used is European Soccer Dataset that contains 8 tables ,, there are relationships among them, to get these tables from sqlite database I maked a connection then wrote query to extract each table and load it in a Data frame.

- Exploratory data analysis section , I have 9 questions to analysis .

4.1- Question 1-what are the best teams over all the seasons?

4.2-Question 2-which 10 Players had the most penalties?

4.3-Question 3-Team attributes lead to most victories?

4.4-Question 4-is there correlation between the number of cards and the outcomes of the match?

4.5-Question 5-what the top 6 players in 2016 ?

4.6-Question 6-How many matches in each season?

4.7-Question 7-How many matches in each league?

4.8-Question 8-show the number of each outcomes for home teams and away teams , Is there a relationship between a team playing at home and winning a match?

4.9-Question 9-Is there a correlation between the number of corners a team has and the number of goals the team scores?

Question 1: what are the best teams over all the seasons?

1-i will get the best team based on the maximum number of goals.

2-used match table to know the number of goals for each team in each season.

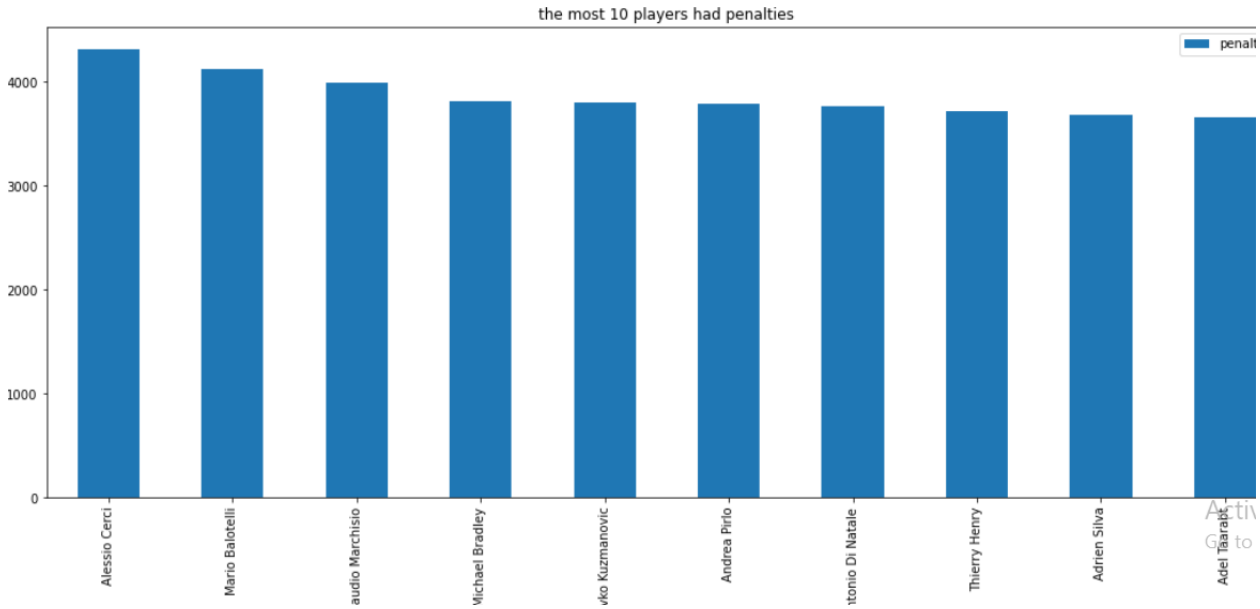
3-used team table to know the names of teams based on id's (the relationship between them).

	season	away_team_api_id	away_team_goal	team_name
0	2008/2009	10281	44	Real Valladolid
1	2009/2010	10281	42	Real Valladolid
2	2010/2011	108893	49	AC Arles-Avignon
3	2011/2012	10269	51	VfB Stuttgart
4	2012/2013	10281	52	Real Valladolid
5	2013/2014	158085	52	FC Arouca
6	2014/2015	274581	53	Royal Excel Mouscron
7	2015/2016	274581	47	Royal Excel Mouscron

Question 2: which 10 Players had the most penalties?

1-used the player_attributes table to know which id_players had the most penalties .

2-used player table to get the players name based on id's (the relationship between player_attr table and player table).



the above plot show the names of 10 players had most penalties over all seasons

the X-axis has names of Players

the Y-axis has numbers of penalties

Question 3:Team attributes lead to most victories?

1- used team attributes table and based on the best teams that i get in question 1 i print the most frequantly attributes of them.

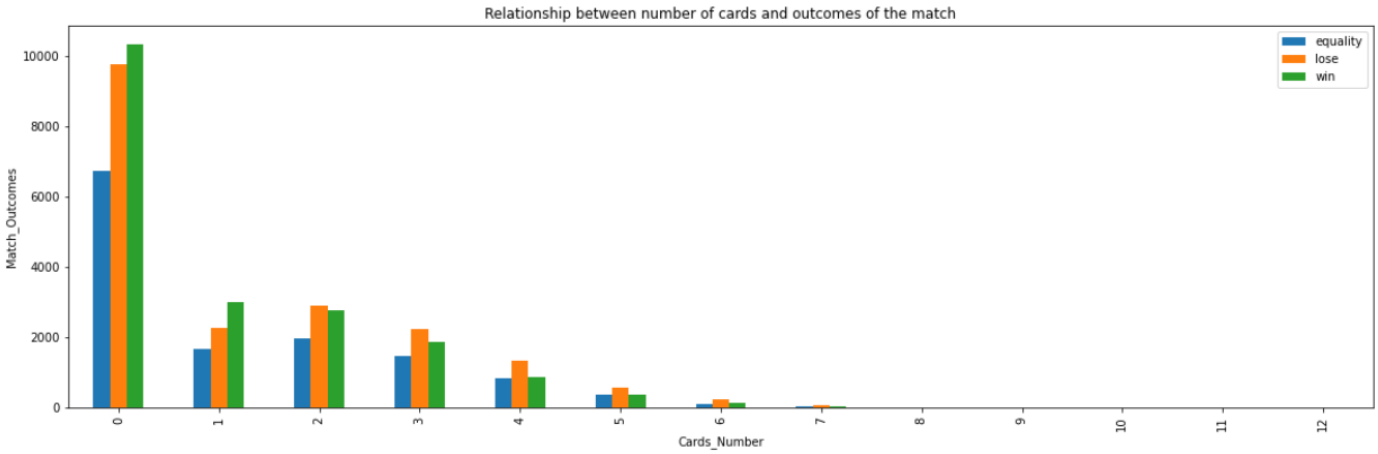
	attributes
buildUpPlaySpeed	55
buildUpPlaySpeedClass	Balanced
buildUpPlayDribbling	40
buildUpPlayDribblingClass	Little
buildUpPlayPassing	54
buildUpPlayPassingClass	Mixed
buildUpPlayPositioningClass	Organised
chanceCreationPassing	58
chanceCreationPassingClass	Normal
chanceCreationCrossing	73
chanceCreationCrossingClass	Lots
chanceCreationShooting	41
chanceCreationShootingClass	Normal
chanceCreationPositioningClass	Organised
defencePressure	51
defencePressureClass	Medium
defenceAggression	50
defenceAggressionClass	Press
defenceTeamWidth	60
defenceTeamWidthClass	Normal
defenceDefenderLineClass	Cover

Question 4:is there correlation between the number of cards and the outcomes of the match ?

1-used match table and classified the outcomes of each team to [win, lose, equality].

2-get the number of matches for each card number and outcome[win, lose ,equality].

3- convert the values of all_outcomes column to columns and their values are the number of matches in each card_number.



the above plot show the relationship between cards number that a team taked in a match and outcomes the team scored in this matchX-axis is Cards_number

Y-axis is Number of matches for eqality ,lose win

conclusion of this plot that there is relationship between cards number that a team taked in a match and outcomes the team scored ,,more number of cards ,, less wining the match :(

less number of cards ,, more chance for wining the match :)

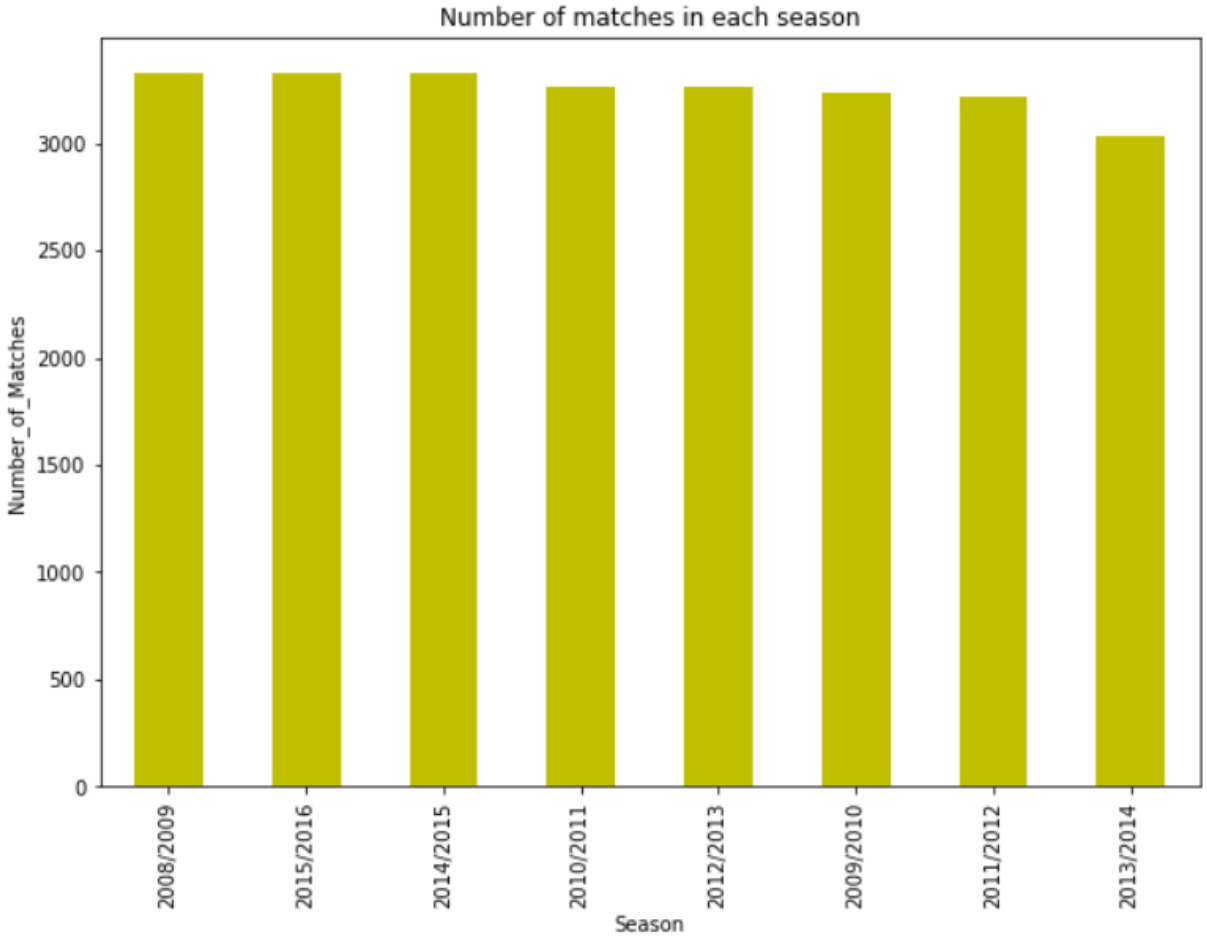
Question 5-what the top 6 players in 2016 ?

1- get the top players based on their avg overall rating in 2016.
2- use player attributes table and convert the date column to datetime to extract the year.
3- filter the table of player attributes in 2016.
4- searching on the name of the 6 top players in 2016 based on ids.

	player_api_id	player_name	overall_rating
0	27299	Manuel Neuer	90.0
1	19533	Neymar	90.0
2	30834	Arjen Robben	89.0
3	37412	Sergio Aguero	88.0
4	36378	Mesut Oezil	88.0
5	107417	Eden Hazard	88.0

6-Question 6-How many matches in each season?

1-get the unique values for season column and count the repetitive values of each season with build in function is called value_counts



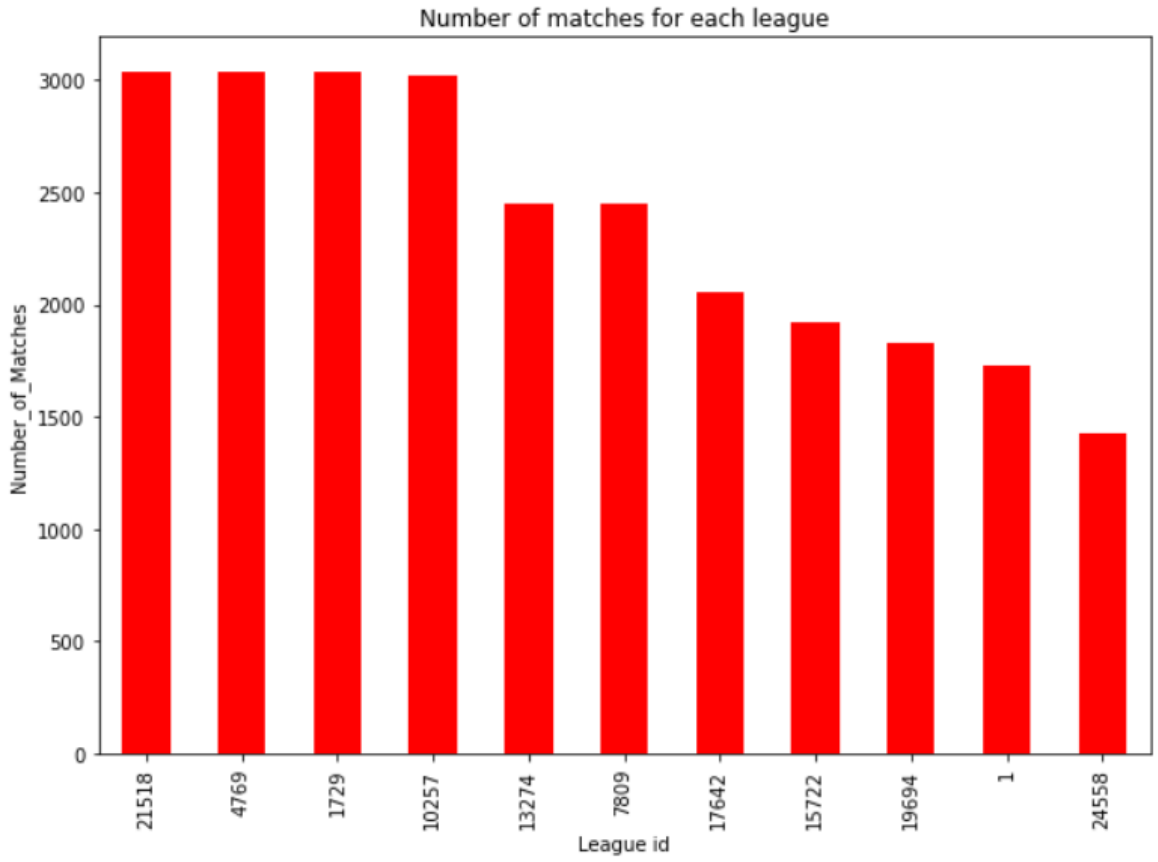
the above bar plot show the number of Matches in each season

X-axis is season.

Y-axis is Number of Matches of each season.

Question 7-How many matches in each league?

1-get the unique values for League column and count the repetitive values of each league with build in function is called value_counts



the above bar plot show the number of Matches in each League

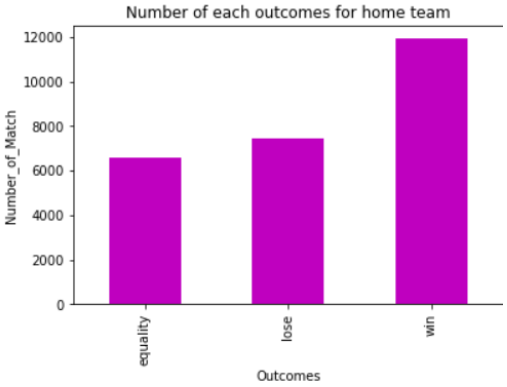
X-axis is season.

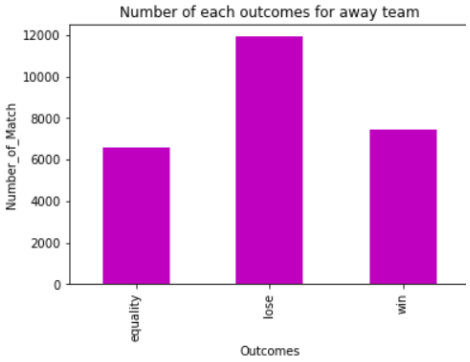
Y-axis is Number of Matches of each League.

Question 8-show the number of each outcomes for home teams and away teams , Is there a relationship between a team playing at home and winning a match?

1-get the unique values for home_outcome, column and count the repetitive values of each outcome with build in function is called value_counts

2- get the unique values for away_outcome, column and count the repetitive values of each outcome with build in function is called value_counts





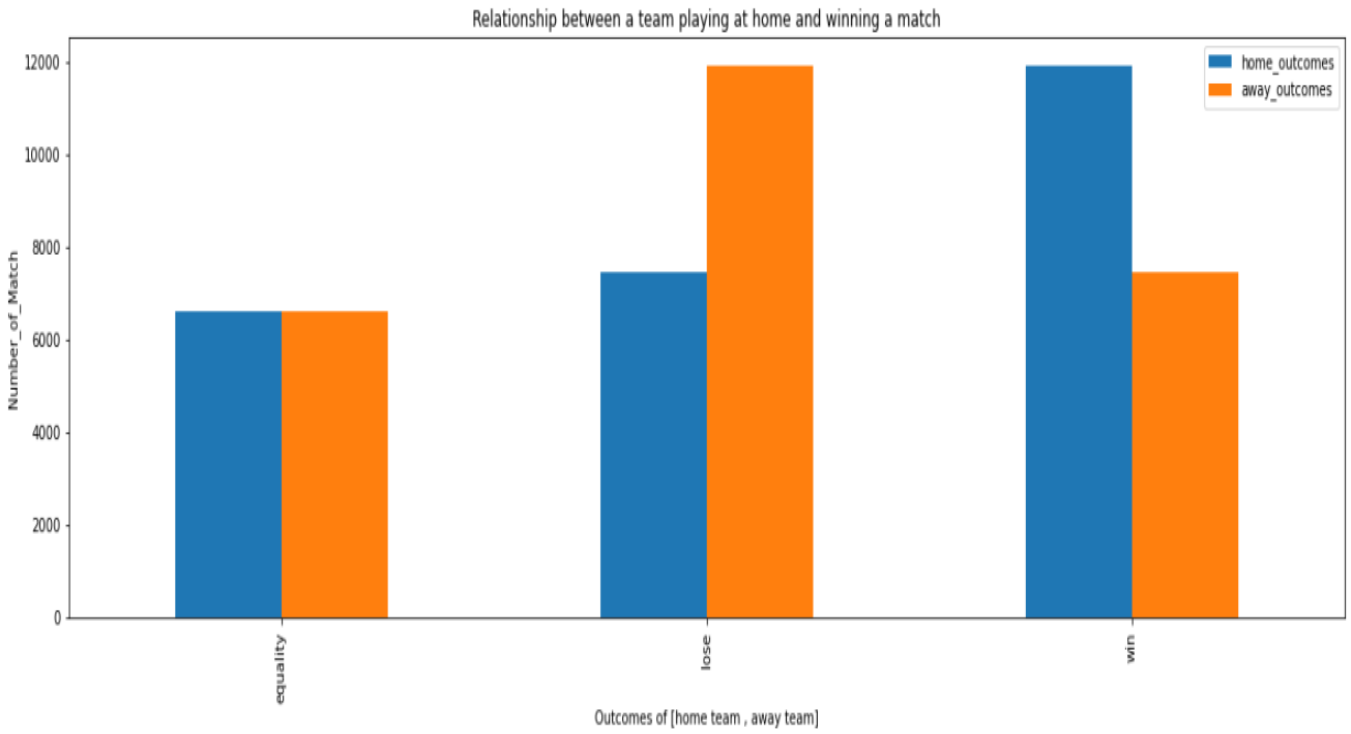
the above two plots show the number of Matches in each Outcome for home,away team

X-axis is outcomes

Y-axis is Number of Matches

1-conclusion that win outcome in most matches for home team :)

2-conclusion that lose outcome in most matches for away team :(



the above plot show the Relationship between a team playing at home and winning a match

X-axis is Outcomes

Y-axis is Number of Matches of each season

conclusion of this plot that there is Relationship between a team playing at home and winning a match,

when the team playing in it's home there is a bigger chance to win the match :)

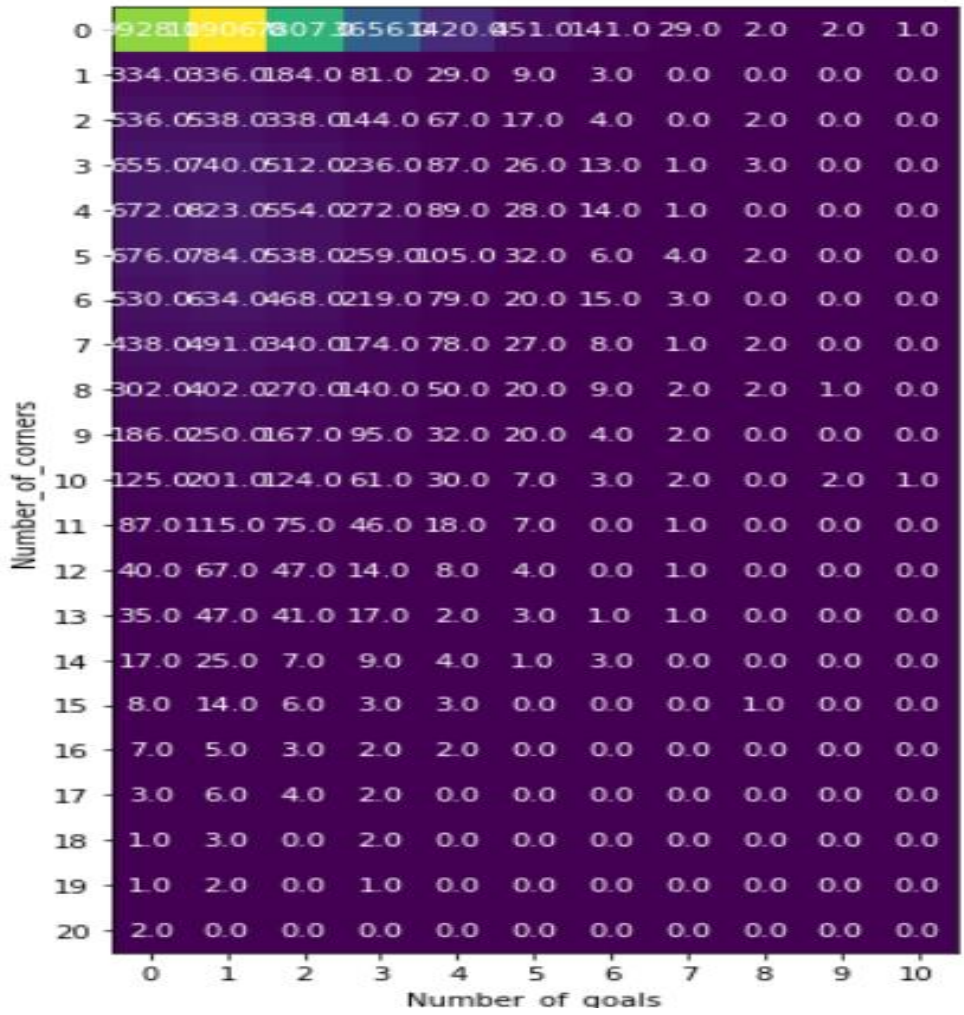
Question 9-Is there a correlation between the number of corners a team has and the number of goals the team scores

1-append 'home_team_goal' and 'away_team_goal' coulmns in one column caleed all_goals

2-append 'home_corner' and 'away_corner' coulmns in one column called all_corners.

***3*- get the 'all_goals' and 'all_corners' in DataFrame is called df_corners and insert column of ones**

4-convert 'all_corners' coulmn to index column and 'all_goals' to columns and it's values the 'number_of_matchs' column



the above Heatmap show the correlation between the number of corners a team has and the number of goals the team scores

X-axis is Number of goals

Y-axis is Number of Corners

conclusion of this plot that there is correlation between the number of corners a team has and the number of goals the team scores,

the less number of corners ,,more number of goals :).

-Data wrangling

cleaning the match table ,, removed the columns that i wouldn’t use and maked parsing for card column for each team and maked parsing for corner column for each team in the match table,then, i removed the sqlitesequence table because i wouldn't use it**.**

-Conclusions

- Limitation

there is hindrance such

1-missing values in Match , Team_attributes Player_attributes Tables

2-data in html form need to scraping

3- match table has alot of columns with Incomprehensible titles.