Investigate A Dataset

Dataset: Soccer Database

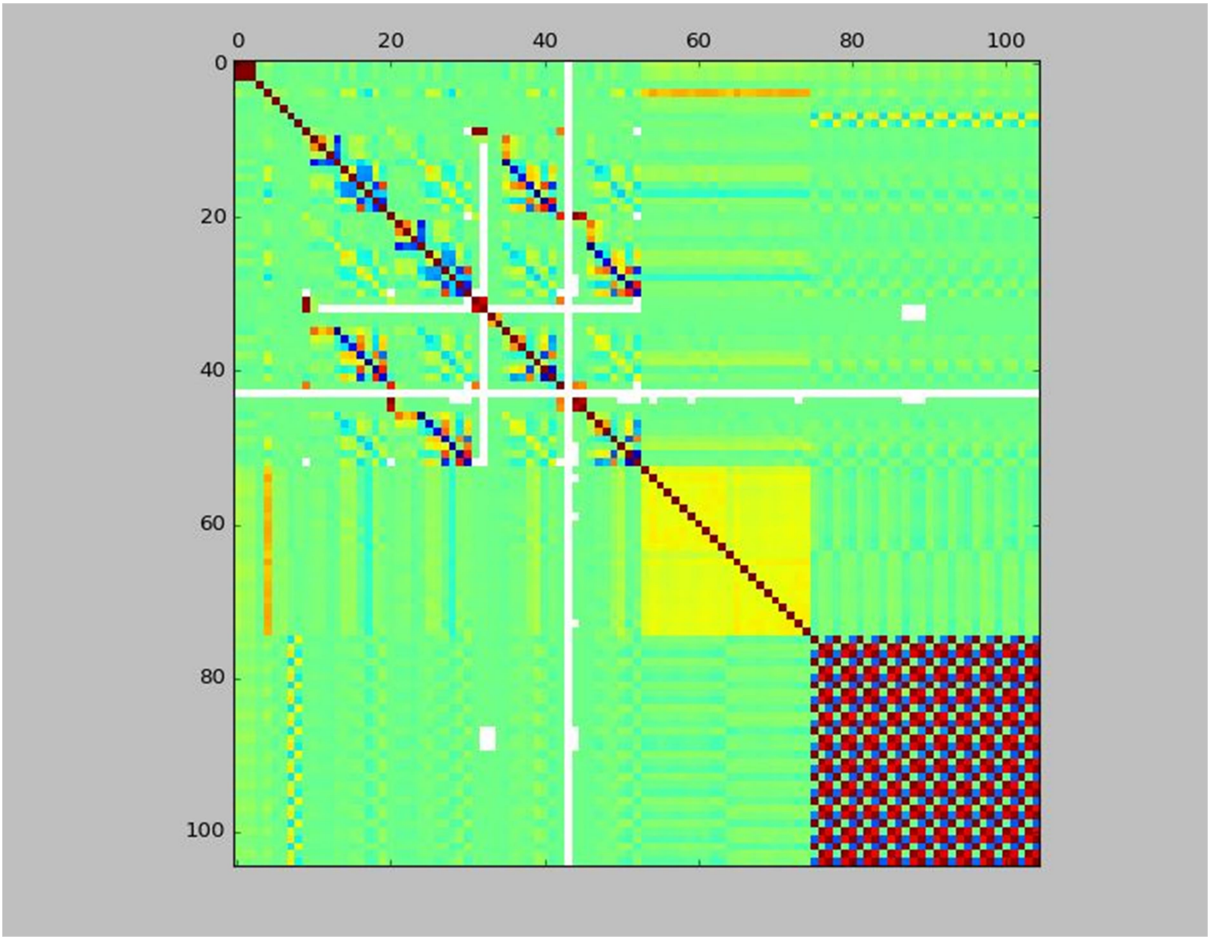
I have mainly tried to find out answers to 3 questions from the dataset which I would be explaining below.

1. Which is the most competitive league?
2. Which is the most attacking team?
3. Where are top teams most likely to score, home or away?

For accessing the dataset I used DB Browser to read the database file with SQL queries and exported each table data onto a csv file.

Data Wrangling

The 'match' table contained lots of highly correlated columns with null values which are not required for my analysis so I removed them.



Also columns from other table which was redundant were also removed.

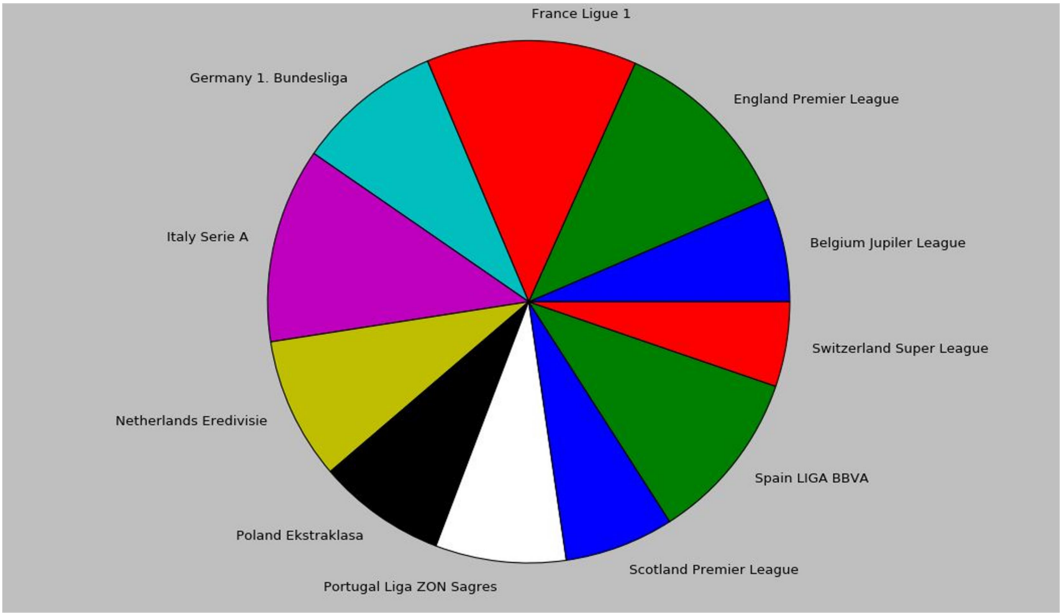
1) Which is the most competitive league?

As most competitive leagues or teams usually play out a draw, I extracted drawn matches from the *'match'* table.

I filtered the matches where home team and the away team have equal goals onto *'draw_matches'*.

Now in order to find the leagues which have the most draws I counted the number of times the particular 'country_id' is occurring in 'draw_matches' and added the count to 'league_most_draw'.

For visualization a pie chart was created on 'league_most_draw'.



Based on the chart the **French Ligue 1** was found out to be the most competitive league followed by 'Italy Serie A' and 'England Premier League'

2) Which is the most attacking team?

Now attacking teams can be those teams who have scored a lot of goals but this can be over a period of time also which may include lots of matches. *(See **Interesting Fact** at the bottom of Document)*

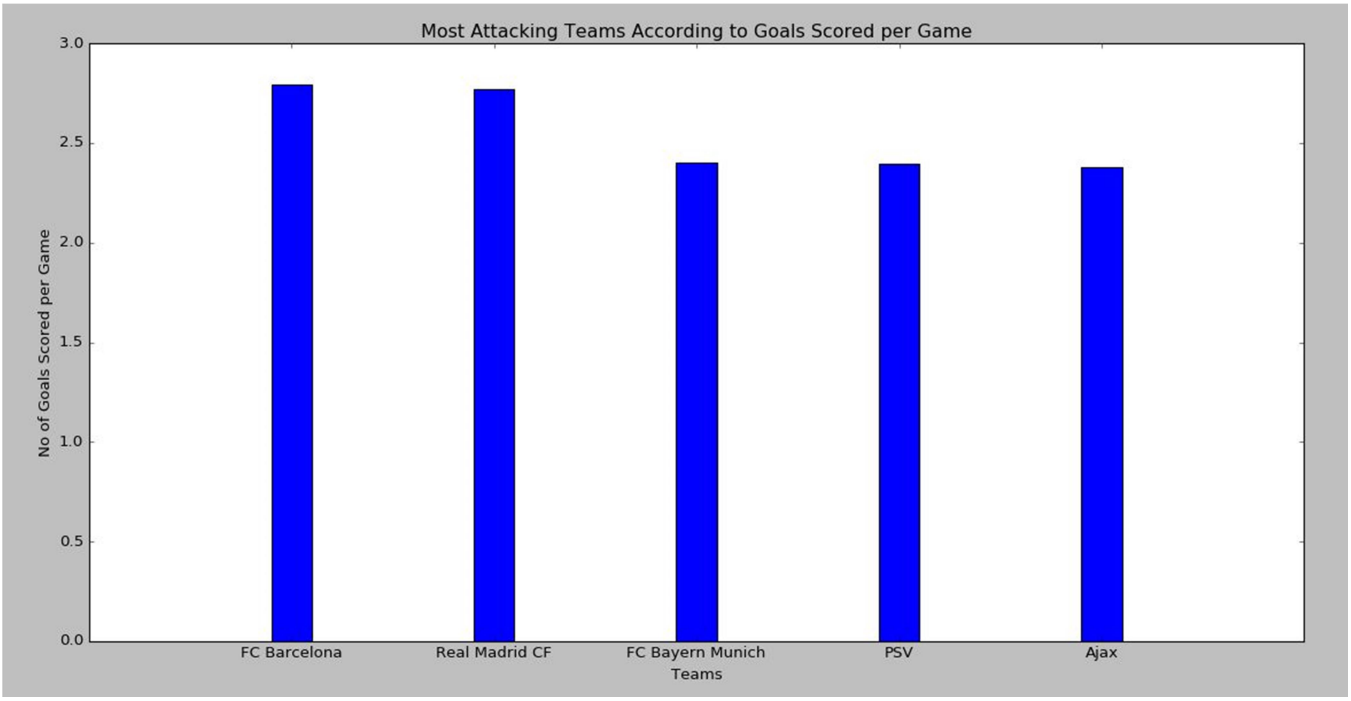
So I considered the most attacking teams to be those who have scored highest number of average goals per game.

I extracted home and away goals scored by each team and stored them in the *'home_goals'* ,*'away_goals'* and *'team_goals'*,.

In order to find the average goals per game, I needed the total matches played by each team also. Similarly I extracted home and away matches count from *'match'* table.

Variable *'total_matches_team'* contains the sum of both home and away matches played by each team. Average goals per game of each team are stored in *'average_goals'*. Top 5 teams with highest goals per game are taken from *'average_goals'* along with their names.

For visualization a bar chart was prepared of these 5 teams.



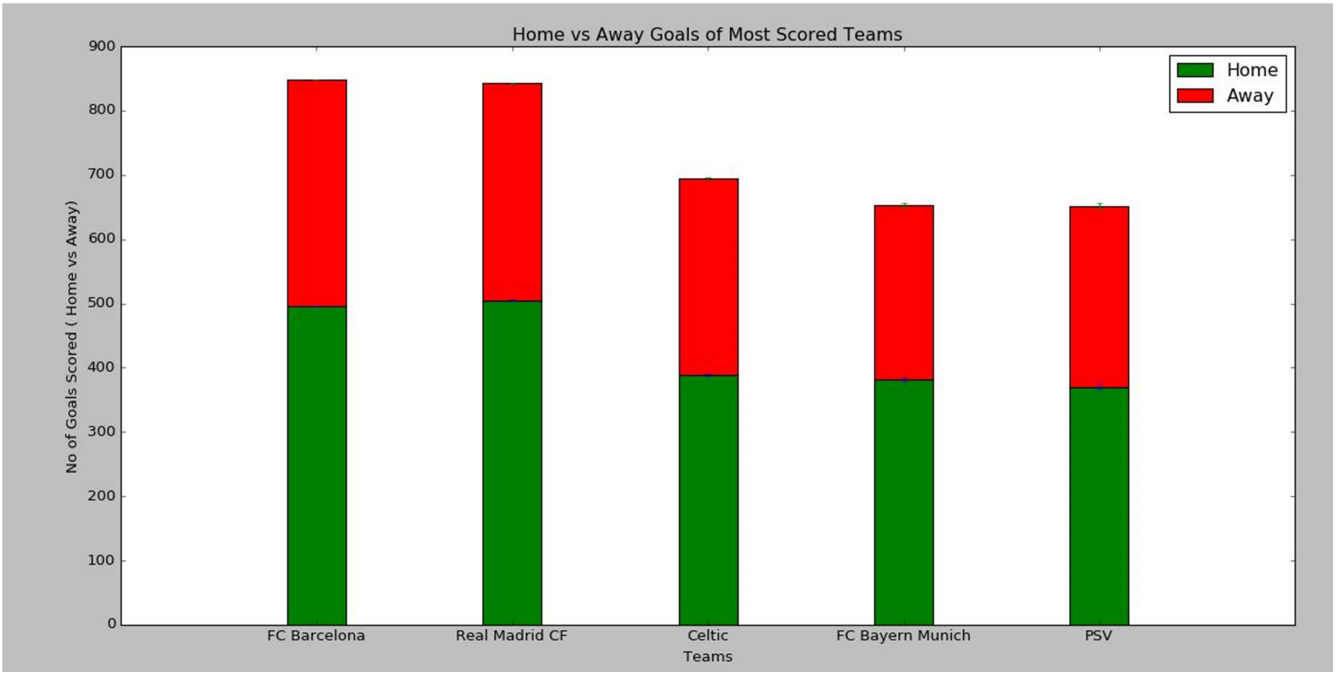
From the above chart it was clear that **Barcelona** were the most attacking team.

3) Where are top teams most likely to score, home or away?

Teams who score the most goals are considered to be the top teams. From '*team_goals'* which I earlier created, top 5 teams with highest goals were extracted onto *'most_goals_overall'* and corresponding teams to *'most_goals_overall_team'*.

Similarly their home and away goals were also put in *'most_goals_overall_home'* and *'most_goals_overall_away'* respectively.

For visualization a stacked bar chart of home and away goals of these top 5 teams were created.



From the chart it was clear that teams playing in **home** always have more chances of scoring than being away.

Interesting Fact: Celtic even though being one of the teams who have scored the most number of goals are not a team who score large number of goals in a game.

Conclusions

After analysing data for the years 2008-2016, it was clear that French Ligue 1 had most number of draws and teams being more competitive. Another point I noticed that scoring more goals does not confirm that the team is more towards the attacking side as your goals per game isn't very high. Also playing in home is an advantage as most goals are scored at home.

Limitations

One limitation that I noticed is that the match data available for each team are not equal. Some teams have played more matches and some less; if equal match data had been present analysis could have been further strengthened.