Presentation:

1. self introduction

2. Domain:

(1) First of all, we are going to introduce the domain of our project, that is the scope of the data.

(2) In this project , we focus on performance of European soccer clubs in past ten years.

(3) To be more specific, we only look at soccer clubs comes from 5 countries:

Spain, England, Germany, Italy and France, since these countries have the so-called 'top-five' soccer

leagues in Europe. And There are 164 teams in total in the data.

(4) For each team we store the data about their performance in domestic league and two European competitions:

the Champions League and the Europa League in each season.

`

3. Q & A:

(1) There are 3 questions in this project. For the sake of coherence of the presentation, we will show the result of

question immediately after presenting the question.

(2) In the first question, we measure a country's performance as a whole. That is, Which country has a better soccer club

performance at European level competitions in the past 10 years? Do Europa League and the Champions League

show similar or different results?

换

(3) Based on the data we have, Spainish soccer clubs dominated Europe in past ten years,

for 100 games they played in the champions league, they won 55 games, which is the highest among all countries.

The second and third best are teams from English Premier League and German Bundesliga.

Italian teams are generally stronger than French teams, but are still worse than other countries.

Also, it is noticable that both the Champions League and Europa leagues show similar results if we think teams

from one country as a whole.

换

(4) In the second question, we will discover some interesting facts about top teams.

That is, Does top-class soccer clubs' performances in their own country's league relate to its performance

in the Champions League in the past 10 years?

换

(5) The answer is indefinite, since we can see different kinds of results in those teams. There are 3 teams that

dominate both continental and domestic competitions. Some teams tend to be more dominate in domestic leagues,

but less competitive in the European stage.

For instance, FC Juventus won every league title in Italy, and has winning rate of 52% in the Champions League.

But FC Arsenal is the opposite of Juventus, it has a lower position in

domestic league, but with a higher winning rate in the Champions League.

换

(6) In the last question, we want to investigate some special factors that influence the performance.

Firstly, we want to investigate Is there a 'Natural Enemy' for clubs in one country to another country

in past 10 years? And Did teams play at home always have a stronger performance than away teams for all the countries?

换

(7) Through the data analysis, we find that Spanish teams always took advantage when competing with Italian teams.

When play in Spain, Spainish teams has a winning rate of 73% On the other hand, when play

in away, Spainish teams still maintained a strong performane, with an unbeaten rate of 72%.

Both of them are larger than the average level

And the location of the game plays an important role in the preformance, we see that in the Champions League, home teams won

about 6 matches in every ten match, and the away team won approxiamte 3 matches in every 10 matches.

换

4. Challenge:

Last but most importantly, when we design our database, there are a few challenges we faced.

One of the challenge is how to store data for domestic leagues.

When we design our database, there are 3 approaches to store the data, the first is to store everything includes

every year, every leagues, every team in a single relation.

That is, we are going to have 3 keys to identify each row: league, year and team Id.

The second approach is for each league, we create a relation. so the key would be season and team.

using this approach, we are going to have 5 relations in total.

The third approach is for every league and each season, we created a relation

if we used this approch, there will be 50 relations in total for this project.

When there are choices, there are trade-off.

First of all, after I learned about design principles, all 3 approchies are valid, and there will be no redundtant data.

The second thing to worry about the effenciy of importing the data, it is obvious that importing 50 datasets are not realistic for this small project, and it is

meaningless, so we give up the last approach.

The third thing to worry about is the complexcity and readability for the data, in my view, storing everything into a single relation leads the project into a mess.

So we give up on the first approach, and thus we chooses the second approach, it balance out effciecy and readability.

However, this decision leads to a slight negative consequence when I want to analysze every league at the same time.

I have to use union operations for several times to merge all domestic leagues.

Most importantly, if we want to further extend our scope to other leagues,

we will hardcode union operation for more times. But this is avoidable if we used the single relation approach.

The second challenge we faced is how to maintain the key of relations. I did not use unique numeric ID for matches.

This reduce my workload, because I have to use data cleaning tools like pandas to give every match a special ID.

In this project, there is no need to store team ID, because we only have a single relation about match results.

But in the future, if we create another relation of detailed information about a match, and then not having numeric ID would be problematic

because there there is no easy way we can link those relations together.

Thus, this approach is bad for future extention.

In conclusion, th biggest lesson we learned in those challeges is that we need to always keep an eye on future extention.

Some hard code may be easy for current situation, but it may lead to serious problem in the future.