ΜΥΕ030 - ΠΡΟΧΩΡΗΜΕΝΑ ΘΕΜΑΤΑ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΕΦΑΡΜΟΓΩΝ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΕΡΓΑΣΙΑ ΓΙΑ ΤΟ ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2024-2025

Χριστόφορος Βασιλάκος – 4861

1 ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

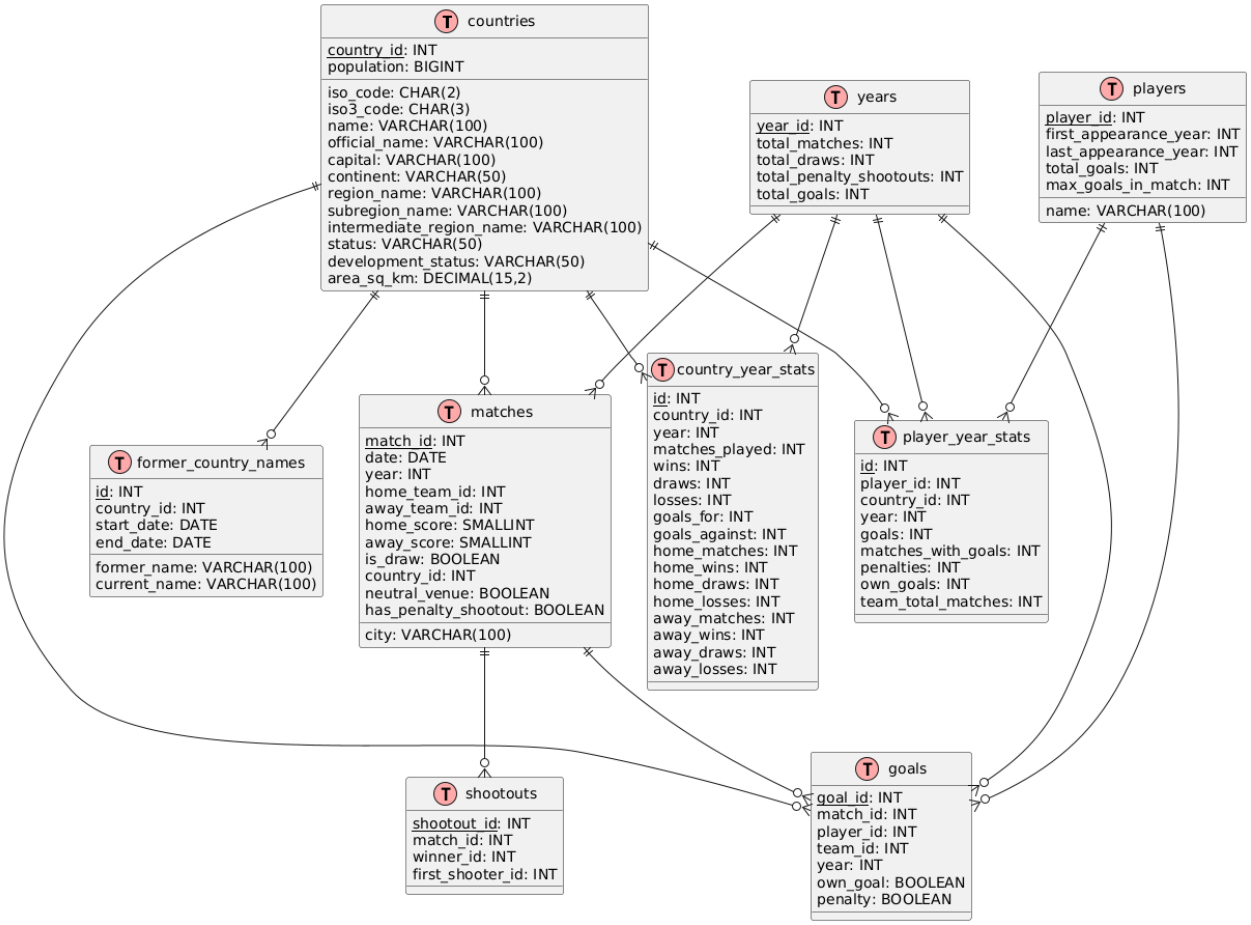


Figure 1: Σχεσιακό σχήμα της βάσης δεδομένων

Ο πίνακας countries περιέχει πληροφορίες για κάθε χώρα, όπως γεωγραφικά και δημογραφικά στοιχεία. Οι μετονομασίες χωρών καταγράφονται στον πίνακα former_country_names. Οι αγώνες αποθηκεύονται στον πίνακα matches, με συνδέσεις προς τις συμμετέχουσες χώρες και το έτος. Τα γκολ καταγράφονται στον πίνακα goals, συνδέοντας παίκτες, αγώνες και ομάδες. Ετήσια στατιστικά στοιχεία για χώρες και παίκτες αποθηκεύονται στους πίνακες country_year_stats και player_year_stats αντίστοιχα. Ο πίνακας shootouts καταγράφει τα πέναλτι.

2 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΛΟΓΙΣΜΙΚΟΥ

2.1 Αρχιτεκτονική και δομή ETL

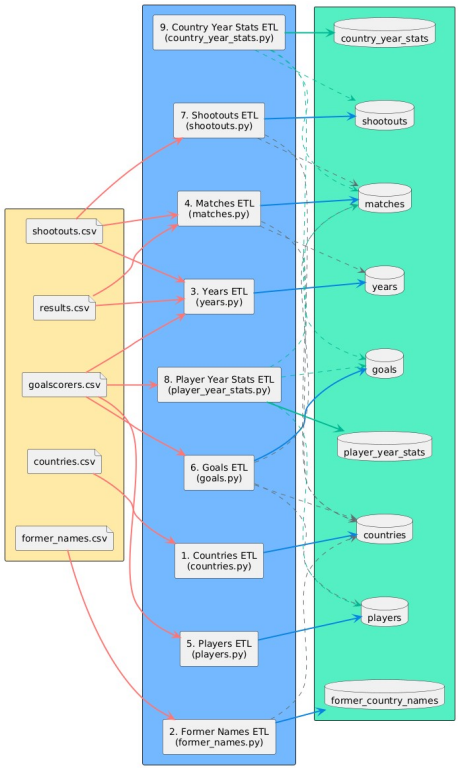


Figure 2: Ολικό διάγραμμα ETL

Μεταξύ των ETL script (μπλε) και database (πράσινου) με διακεκομμένα γκρι βελάκια φαίνονται τα dependacies. Τα ETL script τρέχουν query πάνω σε αυτά.

2.1.1 Countries ETL

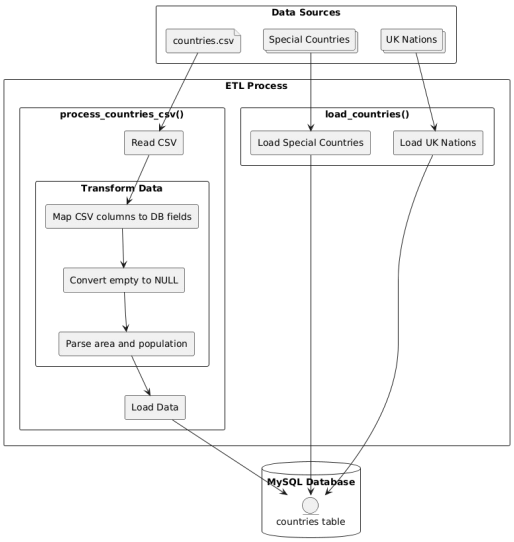


Figure 3: Countries ETL

Ο στόχος του ETL script countries.py είναι να σιγουρέψει ότι τα ονόματα χωρών στο σύστημα, καθώς οι χώρες συχνά εμφανίζονται με διαφορετικές ονομασίες στα δεδομένα που έχουμε για τους αγώνες είναι ακριβές. Το script διαβάζει βασικά δεδομένα χωρών από το CSV αρχείο και επεξεργάζεται ειδικά περιπτώσεις, όπως τα έθνη του Ηνωμένου Βασιλείου (που συχνά εμφανίζονται ως ξεχωριστές οντότητες [Wales, κτλπ]) και χώρες με πολλαπλές ονομασίες (όπως η "China" vs "China PR"). Καθώς και άλλες ειδικές περιπτώσεις όπως χώρες που έχουν περίεργο encoding πχ Aaland Islands.

2.1.2 Former names ETL

Figure 4: Former names ETL

Το ETL script αυτό είναι κυρίως βοηθητικό για τα παρακάτω ETL που χρειάζεται να κάνουν lookup τα ονόματα των χωρών, με αυτόν τον τρόπο θα μπορούμε να επιστρέψουμε τα σωστά id για τα ονόματα αυτά.

2.1.3 years ETL

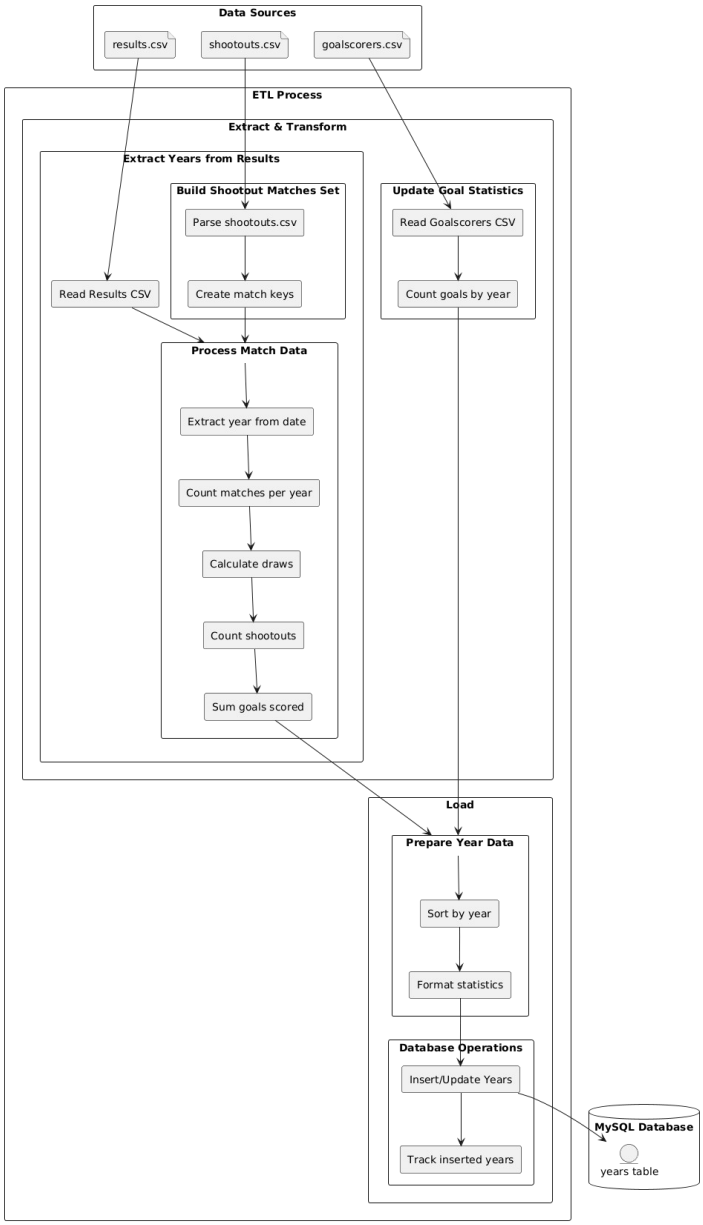


Figure 5: years ETL

Για το table years το id που είναι και το primary key χρησιμοποιώ το actual year, πχ date = 1990-01-01 θα έχει id = 1990. Το id αυτό θα χρησιμοποιηθεί από άλλα table ως foreign key για να γίνουν τα join μεταξύ πολλών πινάκων πιο εύκολα.

2.1.4 matches ETL

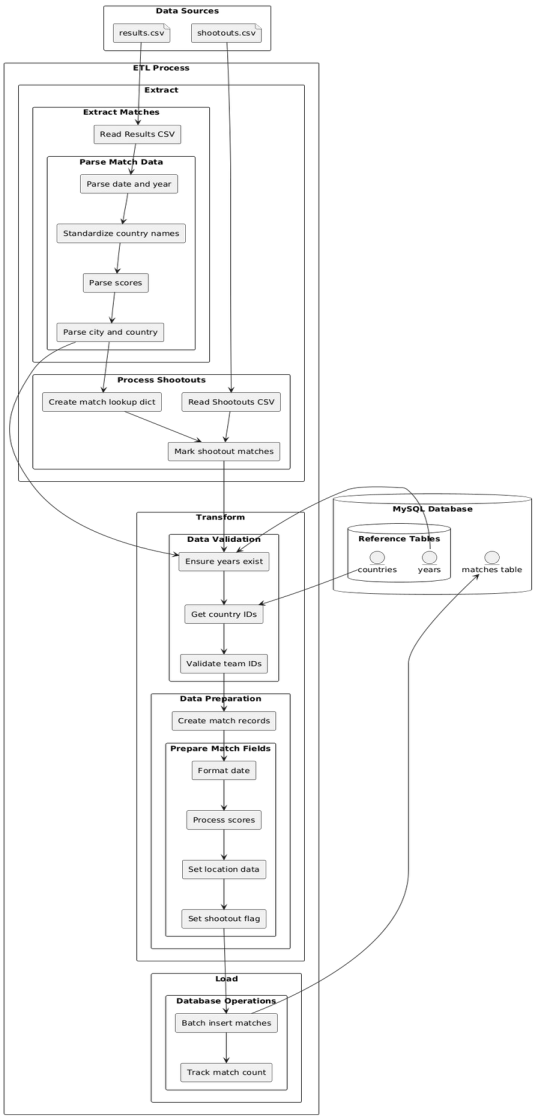


Figure 6: matches ETL

Το ETL script matches.py επεξεργάζεται τα results.csv και το shootouts.csv. Μία από τις κύριες λειτουργίες του είναι να βεβαιώσει ότι υπάρχουν οι χώρες στο countries table καθώς και η χρονία στην οποία έγινε ο αγώνας. Ουσιαστικά παίρνει τα ονόματα των ομάδων τα αντικαταστά με τα id που δείχνουν στο countries table ως foreign keys (το ίδιο ισχύει και για την χρονία).

2.1.5 Players ETL

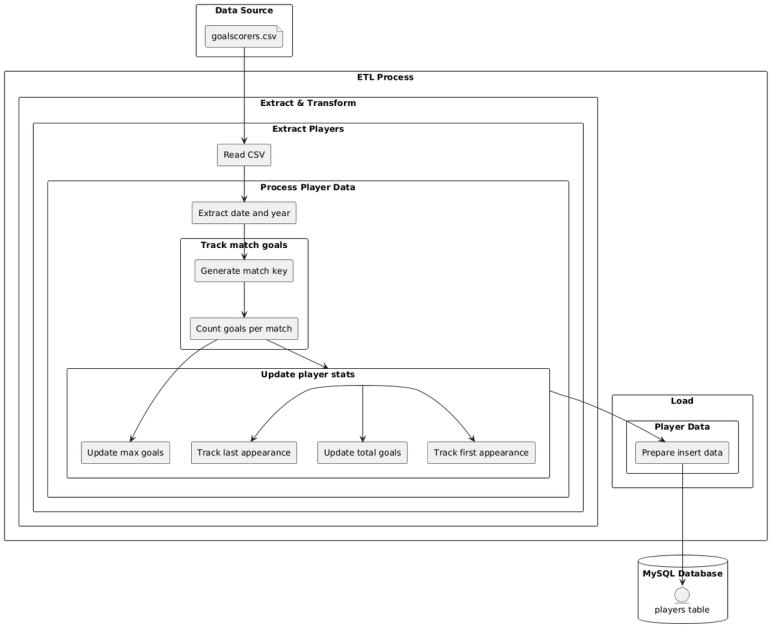


Figure 7: players ETL

Το ETL script players.py κάνει extract πληροφορίες από το goalscorers.csv όπως τον αριθμό των γκολ, πρώτη και τελευταία χρονιά στην οποία έπαιξε ο παίκτης καθώς και τον μέγιστο αριθμό γκολ σε ένα match χρησιμοποιώντας key-value structures (dictionaries python).

2.1.6 goals ETL

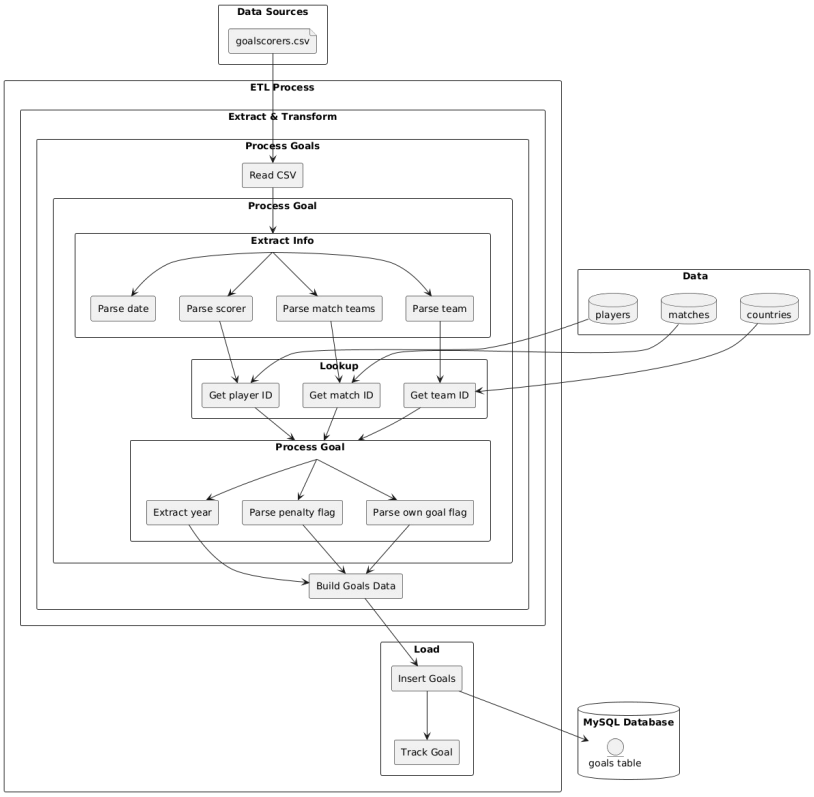


Figure 8: goals ETL

Το ETL script goals.py διαβάζει δεδομένα από το goalscorers.csv για κάθε εγγραφή γίνεται validate κάνοντας lookup τα table matches, players & countries για τα αντίστοιχα foreign keys. Υπάρχει επίσης το foreign key ως προς το table years. Κάνοντας πιο εύκολη την διαδικασία των join για των υπολογισμό των στατιστικών ανά χρονιά.

2.1.7 shootouts ETL

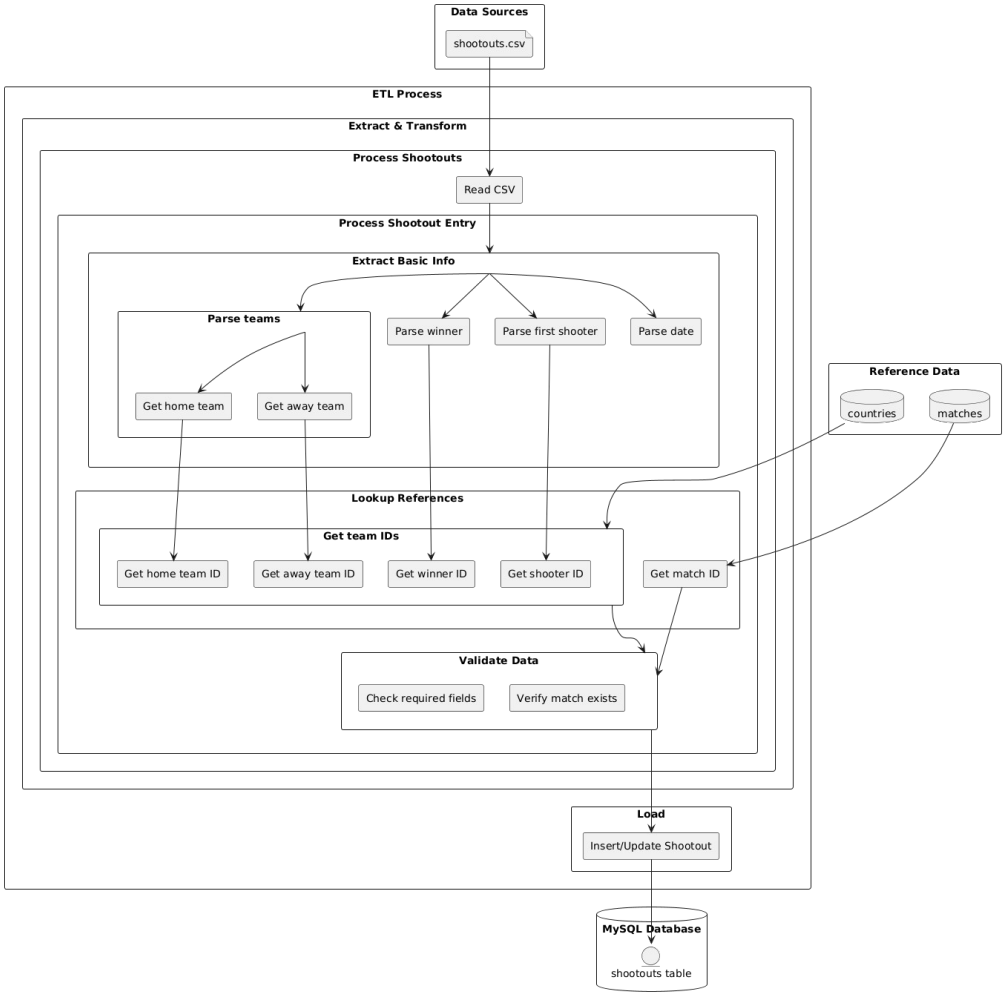


Figure 9: shootouts ETL

Το ETL script shootouts μετατρέπει δεδομένα των πέναλτι από το αρχείο shootouts.csv, για κάθε πέναλτι επαληθεύει ότι ο αγώνας υπάρχει στον πίνακα matches, το ίδιο ισχύει και για την χώρα. Valid εγγραφές γίνονται insert στο shootouts table με foreign key προς το match(match_id) και countries(winner_id, first_shooter_id).

2.1.8 Player year statistics ETL

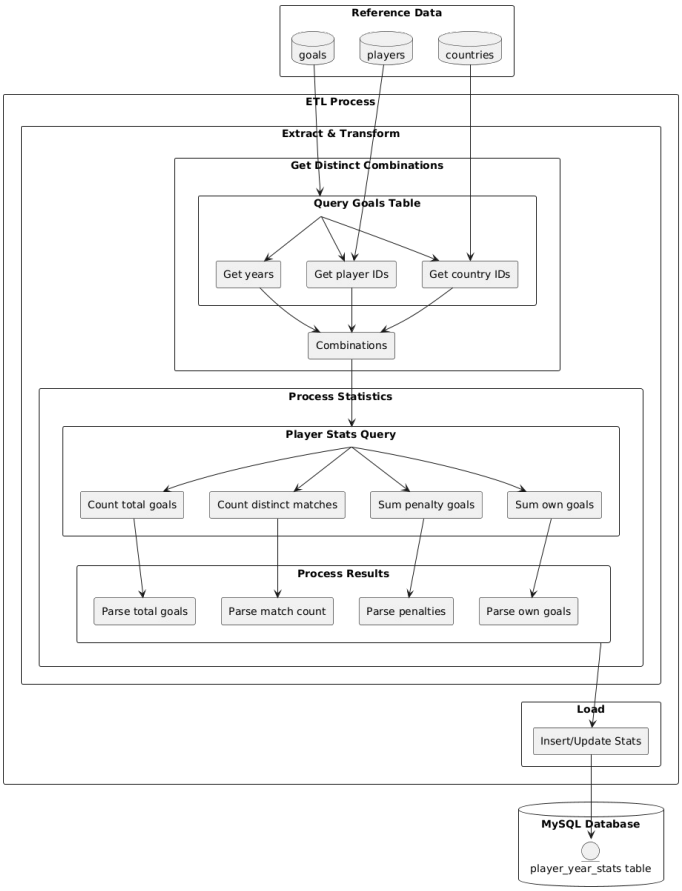


Figure 10: player year stats ETL

Το ETL script player_year_stats δημιουργεί στατιστικά ανά χρονιά για κάθε παίκτη-ομάδα που υπάρχει στα δεδομένα που δημιούργησαν τα προηγούμενα ETL scripts. Στέλνει query στο πίνακα goals για να πάρει όλους τους συνδυασμούς από player_id, team_id (ως country_id) και year. Για κάθε συνδυασμό, υπολογίζει: total goals scored, αριθμό αγώνων που ο παίκτης σκόραρε τουλάχιστον ένα γκολ, κτλπ. Αυτά υπολογίζονται μέσω SQL query που φιλτράρει και γκρουπάρει τα δεδομένα. Τέλος τα δεδομένα εισάγονται στον πίνακα player_year_stats.

2.1.9 Country year statistics ETL

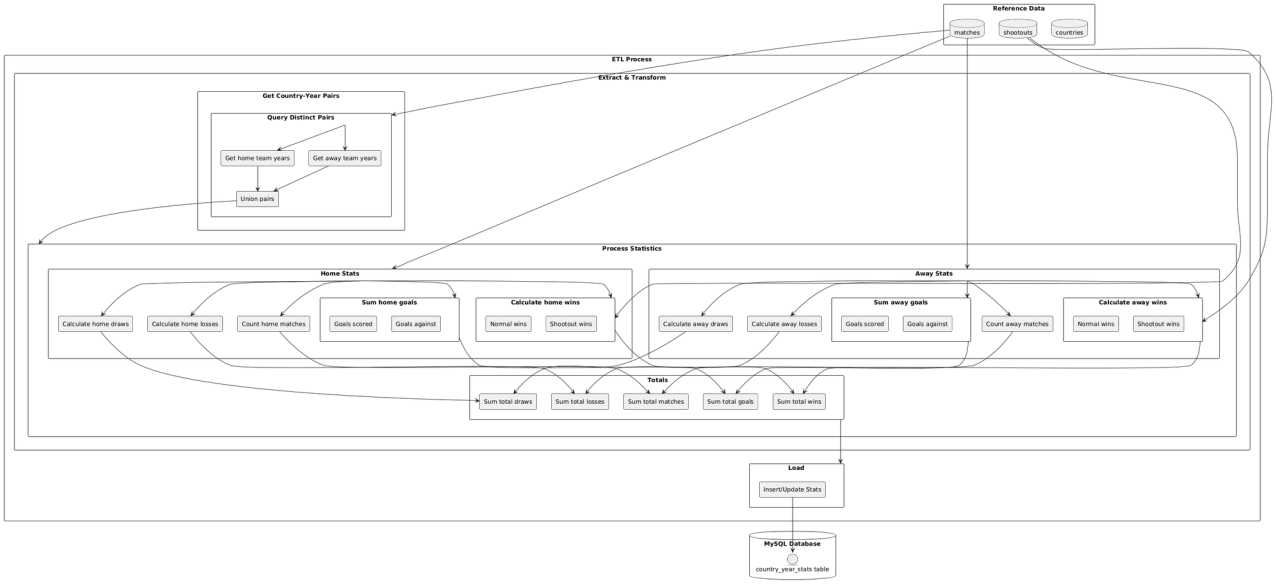


Figure 11: country year stats ETL

Το ETL script country_year_stats παίρνει όλα τα αποτελέσματα των αγώνων για κάθε χώρα σε κάθε χρονιά. Αρχικά βρίσκει όλους τους συνδυασμούς χωρών και ετών από τους αγώνες που έπαιξαν. Για κάθε συνδυασμό, υπολογίζει χωριστά τα στοιχεία για τους εντός και εκτός έδρας αγώνες. Για τους εντός έδρας αγώνες μετράει τον αριθμό αγώνων, πόσες φορές νίκησε (κανονικά ή στα πέναλτι), πόσες φορές έφερε ισοπαλία και πόσες φορές έχασε. Τα ίδια μετράει και για τους εκτός έδρας αγώνες. Για τα πέναλτι κοιτάει ποια ομάδα κέρδισε όταν ο αγώνας τελείωσε ισόπαλος. Στο τέλος προσθέτει όλα τα νούμερα μαζί και τα βάζει στο table country_year_stats.

2.2 Διάγραμμα κεντρικής εφαρμογής

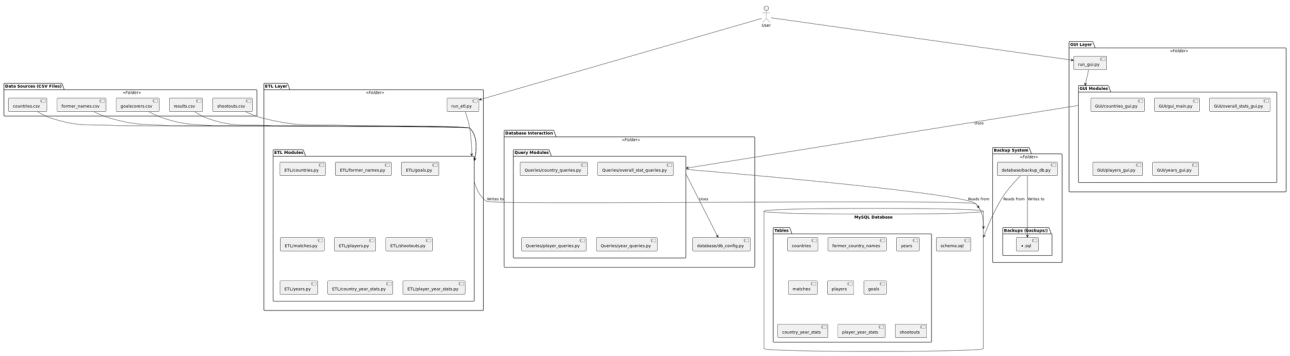
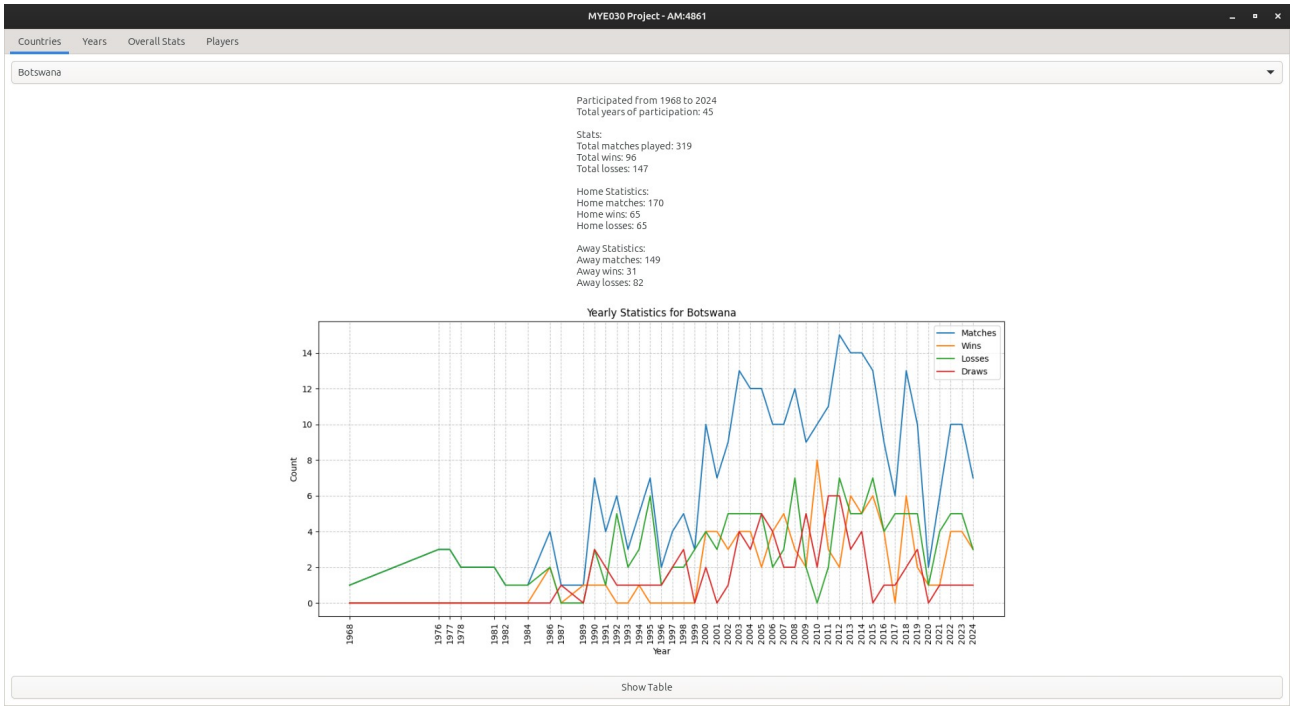
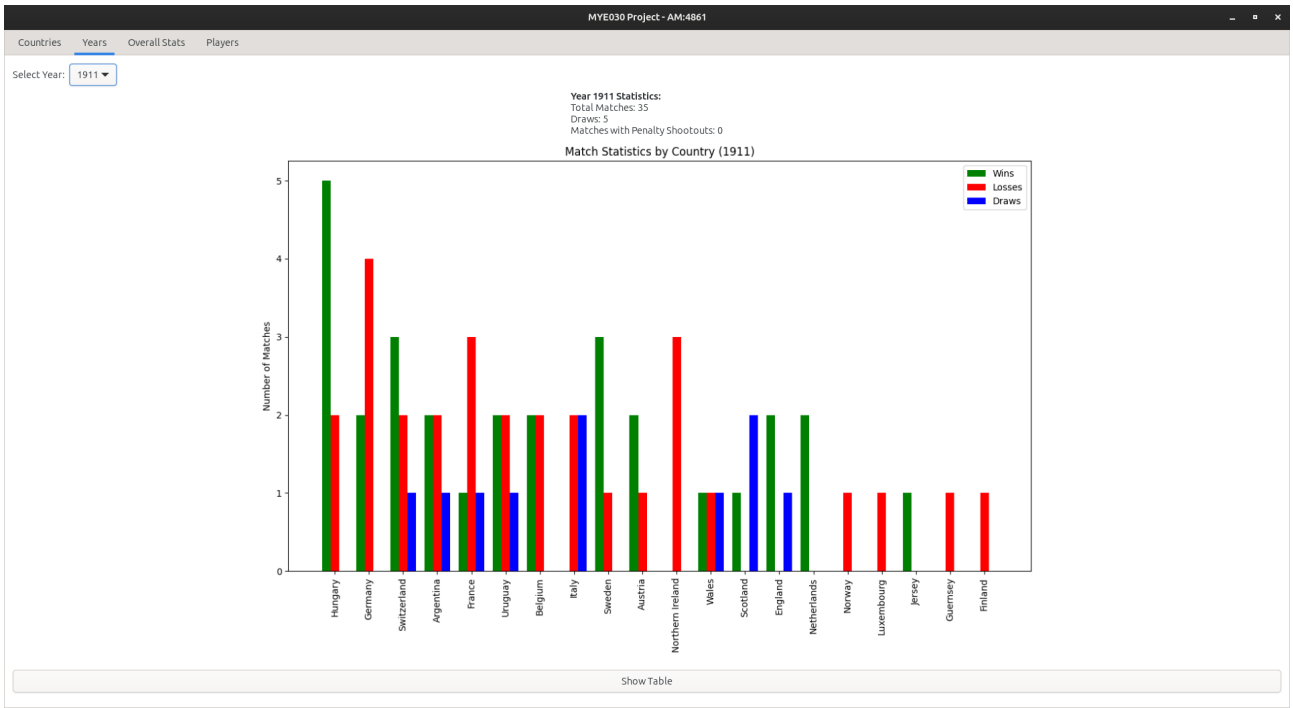


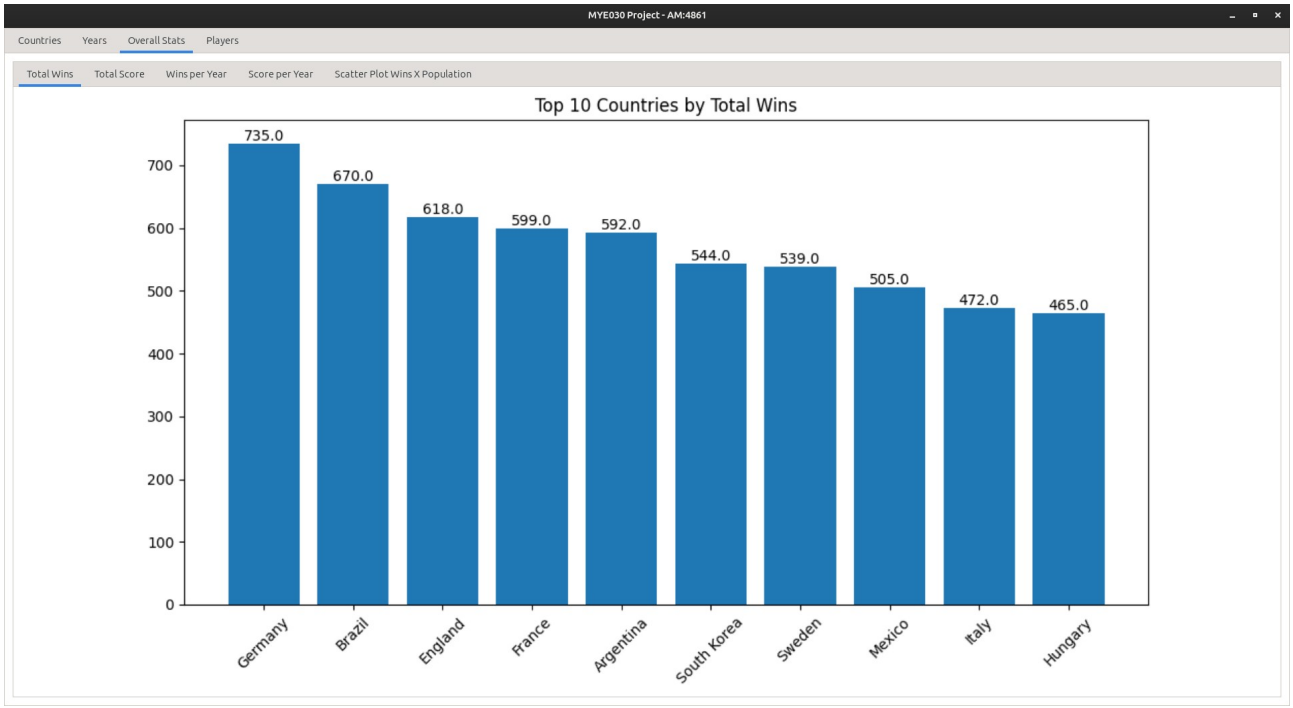
Figure 12: Application Diagram

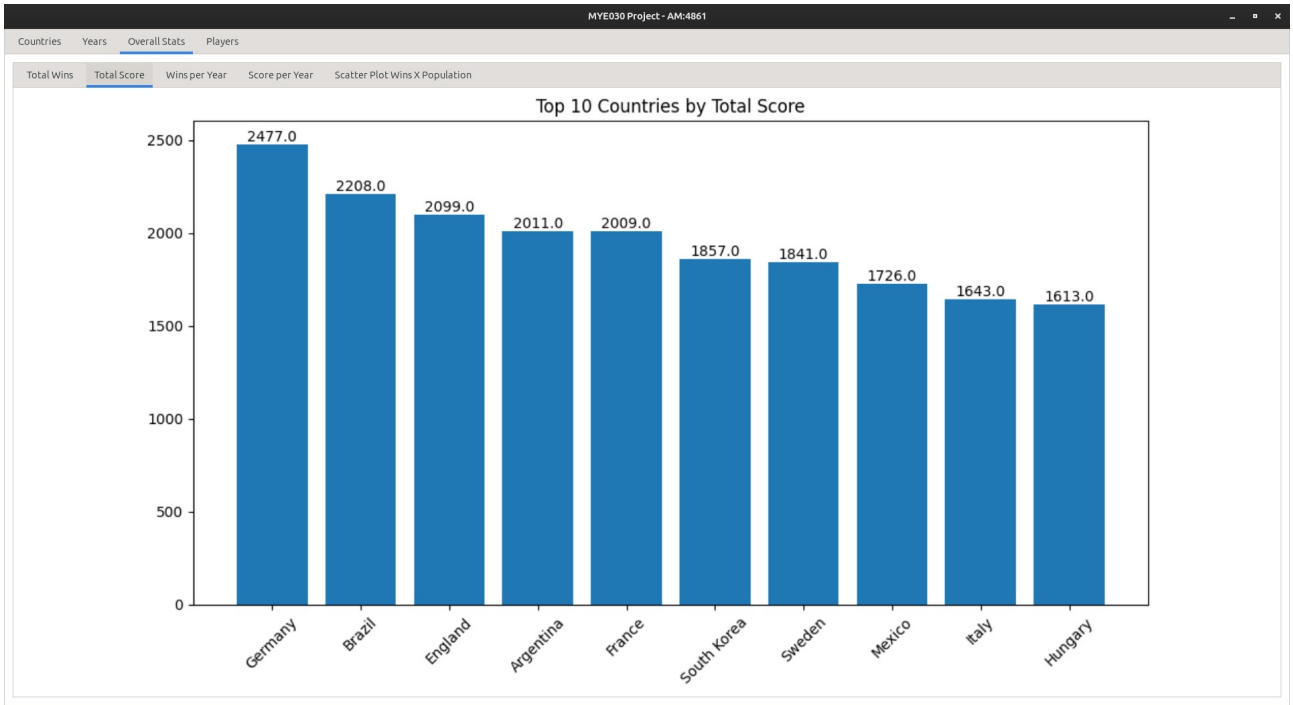
(Επειδή δεν φαίνεται και τόσο καλά θα υπάρχει και στο Github repo)

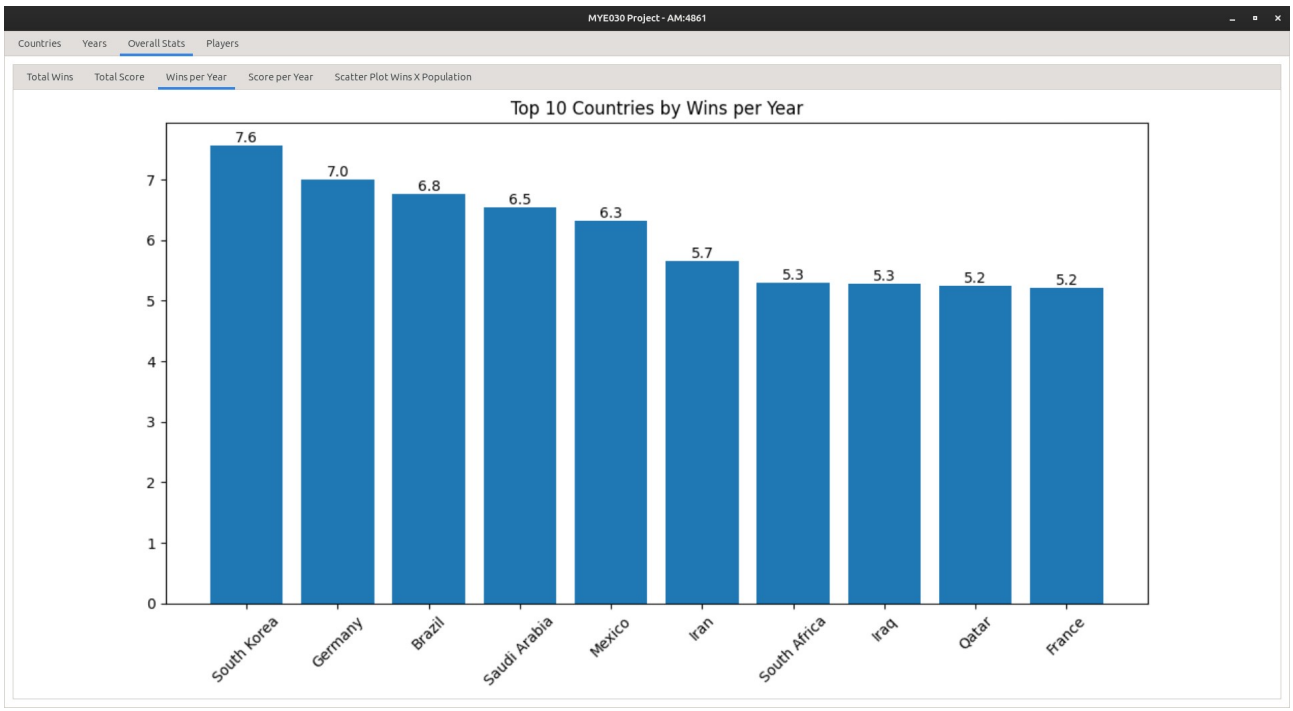
3 Μερικά Screenshot από την εφαρμογή

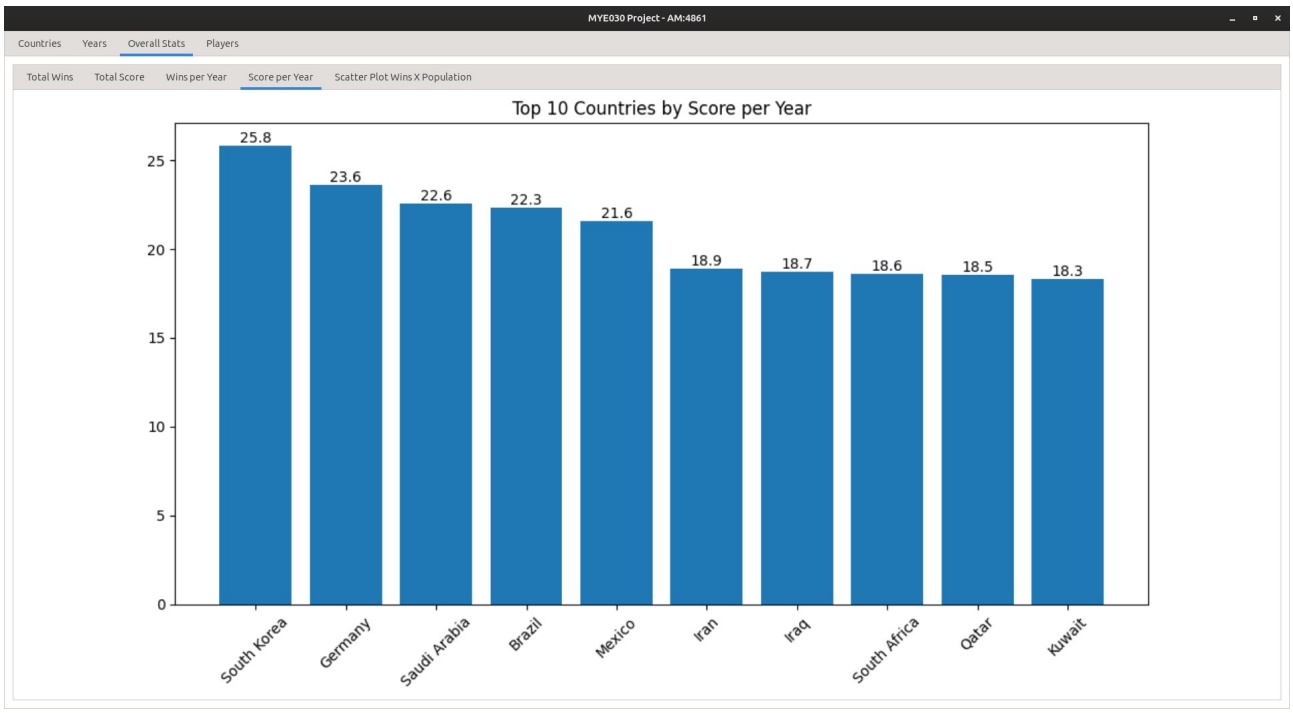


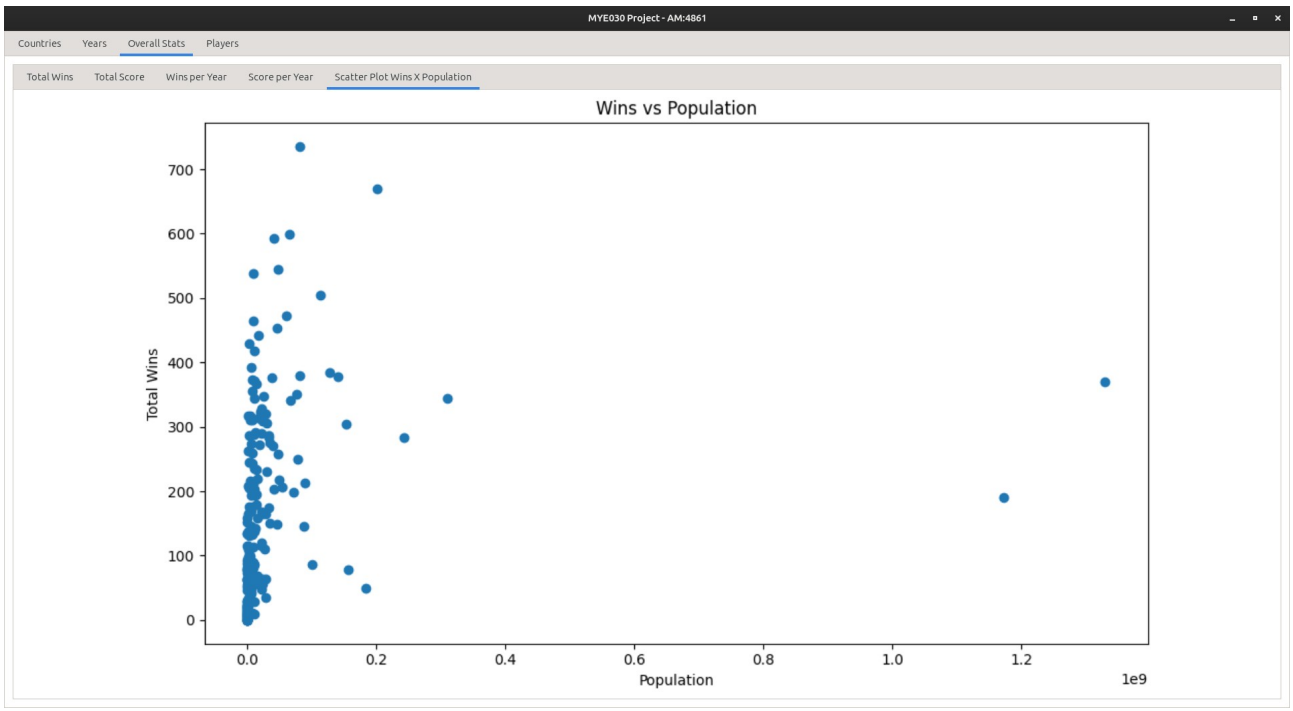


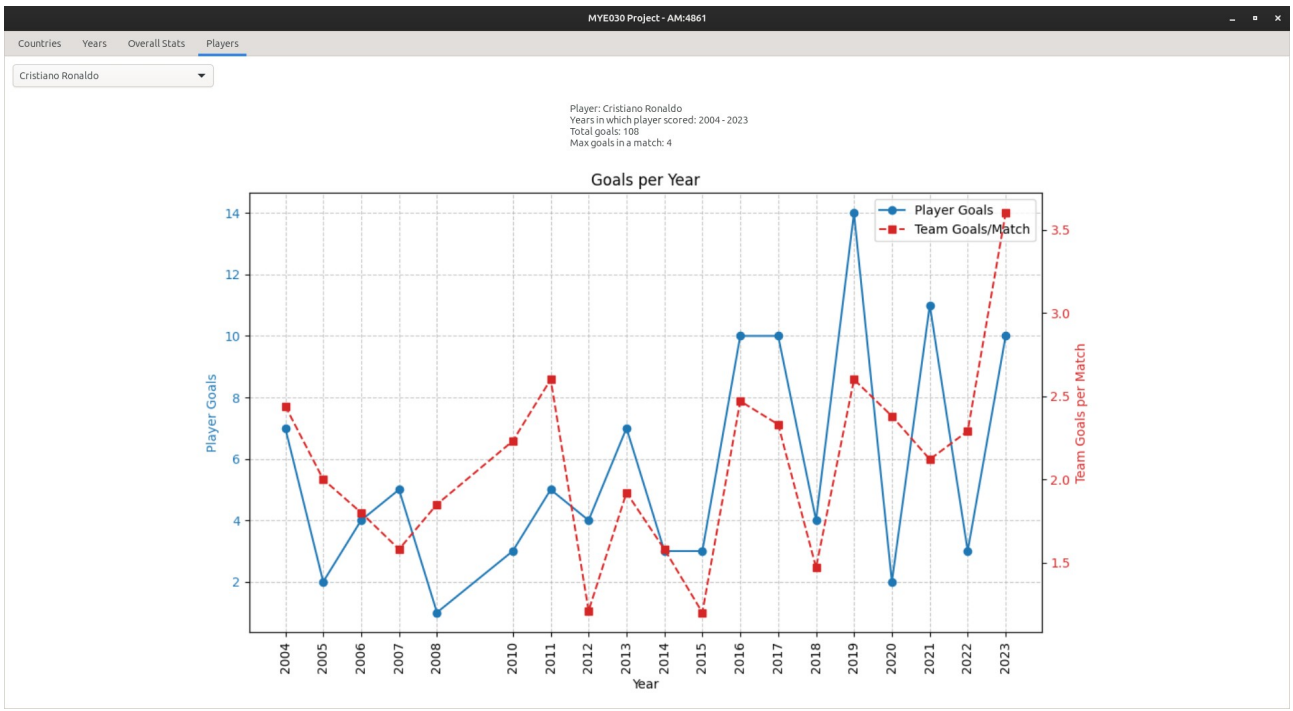

MYE030 Project - AM:4861
Countries
Years
Overall Stats
Players
Total Wins
Total Score
Wins per Year
Score per Year
Scatter Plot Wins X Population
Top 10 Countries by Total Wins
735.0
670.0
618.0
599.0
592.0
544.0
539.0
505.0
472.0
465.0
0
100
200
300
400
500
600
700
Germany
Brazil
England
France
Argentina
South Korea
Sweden
Mexico
Italy
Hungary


MYE030 Project - AM:4861
Countries
Years
Overall Stats
Players
Total Wins
Total Score
Wins per Year
Score per Year
Scatter Plot Wins X Population
Top 10 Countries by Total Score
2477.0
2208.0
2099.0
2011.0
2009.0
1857.0
1841.0
1726.0
1643.0
1613.0
0
500
1000
1500
2000
2500
Germany
Brazil
England
Argentina
France
South Korea
Sweden
Mexico
Italy
Hungary


MYE030 Project - AM:4861
Countries
Years
Overall Stats
Players
Total Wins
Total Score
Wins per Year
Score per Year
Scatter Plot Wins X Population
Top 10 Countries by Wins per Year
7.6
7.0
6.8
6.5
6.3
5.7
5.3
5.3
5.2
5.2
South Korea
Germany
Brazil
Saudi Arabia
Mexico
Iran
South Africa
Iraq
Qatar
France


MYE030 Project - AM:4861
Countries
Years
Overall Stats
Players
Total Wins
Total Score
Wins per Year
Score per Year
Scatter Plot Wins X Population
Top 10 Countries by Score per Year
25.8
23.6
22.6
22.3
21.6
18.9
18.7
18.6
18.5
18.3
South Korea
Germany
Saudi Arabia
Brazil
Mexico
Iran
Iraq
South Africa
Qatar
Kuwait





4 Σχόλια

Τα dependacies που έχει το project και οδηγίες για το πως να τα εγκαταστήσετε ή να τρέξετε την εφαρμογή αν χρειάζεται θα υπάρχουν στον Github repo.