INST327, Section 0202
Project Report
05/09/2022
The Twos : Zain Bhatti, Wonwoo Choi, Jordan Bell, Jorge Argueta, Steven Clark

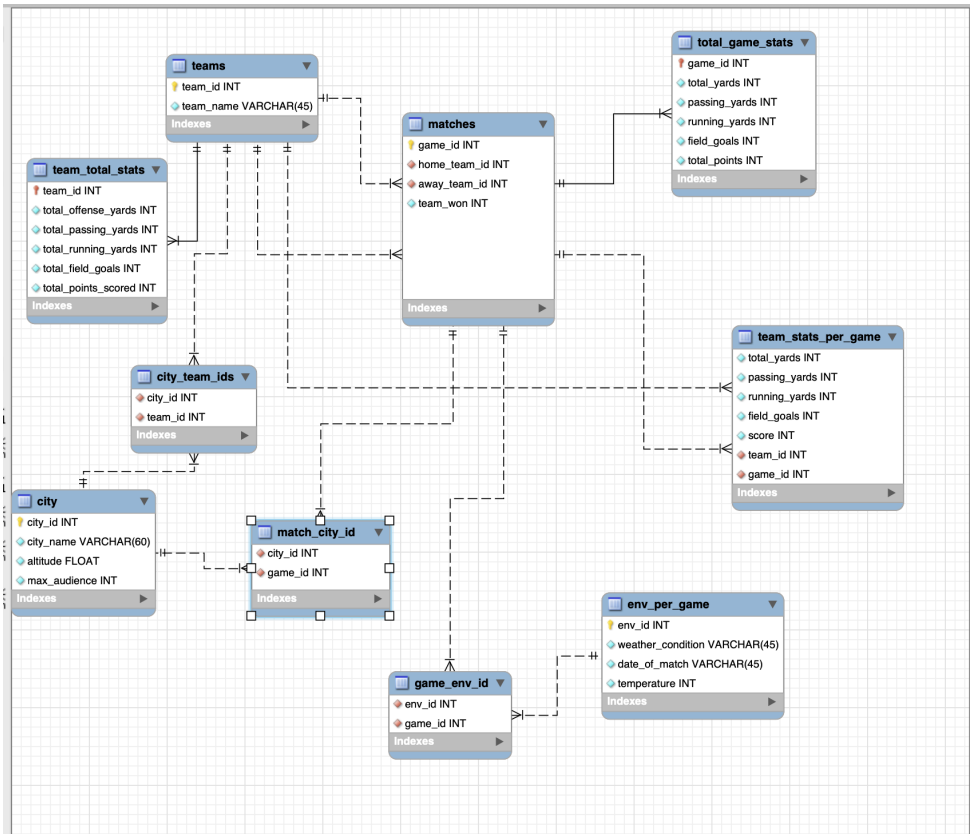
Introduction:

Our team created a database of NFL teams and their match performances in the season of 2021-2022. We recorded yardage by the teams, total stats of the team and the weather condition when the match was played. Our goal was to analyze several factors that could affect how a team plays or performs and if it affects the win rate or not. By having this database, teams will be able to analyze what tactics they should play when they are in a certain situation.

Database Description:

Our database is constructed around the various teams in the NFL and their game to game performance. Many supplementary aspects of each game are included such as the weather, location of play, and audience of each stadium. This serves as a tool to analyze how each team, or both teams combined perform under many different conditions. The database consists of 7 tables and 3 linking tables. All 32 NFL teams are included in this database and 95 games were used in the database (each team's first home and away games, as well as their last home and away games).

Logical Design (include a PNG of your ERD in this section):



We wanted to create a database that allowed users to quickly access information related to NFL games, and make deductions based on the data they are presented. With that in mind, we created the design with ten tables that were somewhat dependent and independent on each other. This allowed the users to take in the database not too complexly and heavily.

Physical Design:

Our goal of creating this database is to analyze what affects the win rates of the teams and the tactics of teams. By analyzing this, teams will be able to get a grasp of what tactics winning teams played in a certain situation like a rainy day or days with high temperature. Data that we retrieved was from ESPN that shaped our physical design. Then we considered what factors could affect game results. For example, weather might have affected the game plans and playing home could have affected players' conditions and how they perform. Therefore we extended from the retrieved database and created tables such as env_per_game and city. Because we needed stats per each game in order to analyze the game, we added tables that showed stats per each game, stats per teams and total stats of the teams: matches, team_total_stats, team_stats_per_game, env_per_game and total_game_stats.

Sample Data:

Our data for game statistics was retrieved from ESPN's website, and includes running yards, passing yards, total yards, field goals, and score. We also used https://www.weatherstem.com/nfl?stadium=5 as a resource for retrieving the temperature a game was played at, as well as the weather conditions at the stadium. We collected relevant data from every team during their first and last home/away game. Using a shared spreadsheet, we entered values for our individually assigned teams and imported the data into MySQL.

Below is a sample of the data present in our team_stats_per_game table.

total_yards	passing_yards	running_yards	field_goals	score	team_id	game_id
260	261	-1	2	6	1	100
434	136	298	1	32	23	100
341	259	82	2	48	30	101
348	295	53	1	25	1	101
257	195	62	2	20	1	102
369	174	195	3	30	19	102
351	118	233	0	29	4	103
265	169	96	1	15	1	103
491	409	82	2	27	22	104
406	217	189	2	33	2	104
481	230	251	1	36	2	105
405	343	62	0	35	14	105

Image showing information from the team_stats_per_game table.

Views / Queries:

Query Names	Req. A	Req. B	Req. C	Req. D	Req. E
altitude_performance	x	x		x	
Games_won_and_Total_points			x		x
home_performance	x	x	x	x	
bad_weather_games_performance	x	x		x	
temperature_performance	x	x			

altitude_performance: This view shows the total combined performance of both teams at different altitudes to see if altitude affects how teams perform. It only looks at games where there is no inclement weather and sorts by total yards.

Games_won_and_Total_points: This will show how many points the teams scored in all of the games stored as well as how many of those games they won. This really helps evaluate the fact that just because a team has the most points, does not mean that they won the most games.

However, there is a strong correlation between the number of points and the number of games won.

home_performance: Creates a view that displays the number of home games won with maximum audiences

bad_weather_games_performance: This view shows the yardage per game but only for games that we considered to be played in "bad weather". Weather scenarios such as heavy rain, light snow, heavy snow, and some more were considered to be bad weather games. The view showed us that in most cases, the bad weather had a huge impact on the number of total yards gained that game and this was not too shocking.

temperature_performace: This query serves to answer whether high temperatures affect performance with respect to total yards. Information returned shows the weather, date, temperature, total_yards, and game_id.

Changes from original design:

We modified and made some changes on some parts of our original plans to produce our final result. One of the biggest changes that we made was reducing our sample size. First, it was impossible to manually input all of the games played by 32 teams. We decided to reduce the number of games to 4; first home and away games and last home and away games. Another major change that we needed was the ERD and tables. When we first proposed our project plan, we did not have an environment table that kept track of weathers and our ERD was not perfect. We applied normalization and adjusted the ERD to make it perfect. Some of the foreign keys were removed such as city_id from the teams table and we made a linking table with team_id and city_id. We also unfortunately had to change the game dates from DATETIME to a VARCHAR variable because it was not converting dates properly and there was some inconsistency in how dates were input into our initial data sheets.

Database Ethics Considerations:

The nature of our database is inherently unbiased and therefore should not have any ethical issues related to exclusion. However, we did consider whether the data we are gathering is fair use. Weather information can come from a variety of outlets, and we were not sure if said outlets held rights to that data. Additionally, since the NFL and ESPN are private entities, we did not want to infringe on what might be considered intellectual property. Upon further

research, we found that all the values we would be considering for our database do in fact fall under fair use. As a result, we focused on making sure the design of the database did not have any kind of personal bias hardcoded. An example could be entering data from games in which the programmers favorite team did well, and excluding games in which they did poorly.

Lessons Learned:

One of the lessons we learned and we agreed was that we need a reliable source that is accurate and exact for the analysis to be accurate. We changed from doing every game of the season to only using 4 games per team. These for games were their first and last home and away games. Since we are dealing with a sample data of 4 games for each team it cannot be said that the result is precise. However, manually inputting hundreds of games would take far too much time. Another lesson we learned was that communication is very important for a project to be done quickly and effectively. Difficulty in consistent meeting times between group members lead to an increased need to effectively communicate to members what needed to be done and by when. Additionally, sometimes when splitting work it can lead to inconsistency in the data that we input, or the occasional duplicate data that was hard to detect.

Potential Future Work:

While our data does a great job in analyzing the general pattern in the teams performance as the season goes through major milestones, it could do a better job in identifying smaller trends. The current status of our database can be described the same way as integrating a line based only on 3 points and the more points, in our case games, we add the more accurate it will be in identifying the integral, in our case the trend in team’s performance, at a specific period of time. Additionally, our database would also benefit in adding more variables that affect the team’s performance. Some examples of more variables are time off between games, wind speed, recovery from injuries, emotional state, timezone and many more. Due to having developed a solid and general structure in the creation of the database, the potential future work that remains is essentially, adding more variables and specifics to analyze with more details how they affect the team's performance.