Malcolm Darukhanawalla

Phase 3 Report

1. **Dataset to Table Conversion**

For this project, I split up the original csv dataset into six different tables. The dataset was split into multiple tables to follow the principles of database normalization, ensuring the data is well-organized, avoids redundancy, and maintains data integrity. I’ve listed each of the tables as well as some example SQL scripts used to create the tables below:

Tables:

1. Games
2. Teams
3. Scores
4. Betting Lines
5. Betting Results
6. Second Half Lines

Sample SQL Script for Games and Betting Lines:

A screenshot of a computer code

Description automatically generated

A computer code with text

Description automatically generated

1. **Challenges/Solutions when Importing Data**
2. Learning how to upload data from csv into table:
   1. Challenge: Initially, I was unfamiliar with the process of uploading data from CSV files into MySQL tables. Identifying an appropriate method to import large datasets was the first obstacle.
   2. Solution: After exploring different options, I discovered the ‘load data infile’ function, which allowed for efficient uploading of data from CSV files. I learned how to use this function, including specifying the file path, delimiter, and other options to correctly align the CSV columns with the table structure.
3. Handling the secure file privilege option
   1. Challenge: I encountered an error indicating that MySQL was running with the secure-file-private option. This restriction meant that CSV files had to be placed in a specific directory defined by MySQL.
   2. Solution: Using the ‘show variables’ statement for secure-file-private, I identified the exact secure directory. This directory was hidden and initially hard to locate, but I managed to navigate to it, place my CSV file there, and proceed with the data import to my tables.
4. Row truncation due to column mismatches:
   1. Challenge: After resolving the secure directory issue, I then faced the “row was truncated” error when trying to import data and it specified it contained more data than there were input columns. This occurred because my original CSV file had more columns than needed for a particular table.
   2. Solution: I split the original CSV file into multiple smaller files, each corresponding to the exact columns required for each table. I ensured the order of columns in the CSV matched the table structure, which resolved the truncation issue.
5. Incorrect date formatting:
   1. Challenge: While populating the Games table, I encountered an error related to an incorrect date format. My CSV file contained dates in the mm/dd/yyyy format, which MySQL does not accept as a standard format.
   2. Solution: I reformatted the date column in the CSV to the MySQL standard format of yyyy-mm-dd using other tools in excel. This modification allowed the dates to be correctly recognized and imported in the appropriate table.
6. Foreign key constraint errors:
   1. Challenge: A significant issue arose when trying to populate tables containing foreign keys, such as the Scores and Betting Lines as well as other tables. Even though I had the correct csv and formatting, I repeatedly kept getting an error that kept me from adding or updating any child rows to these tables due to the failure of foreign key constraints.
   2. Solution: I realized that the foreign key values needed to be present in both the referencing table and the CSV being imported. To resolve this, I added the foreign key columns and their corresponding values to the relevant CSV files. Even though the foreign key values were generated in another table, including them in the CSVs for dependent tables resolved the issue.
7. **Data dictionary for each table**
   * + 1. Games

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Constraint |
| game\_id | Int | Primary key |
| season | Year | Not null |
| date | Date | Not null |
| regular | Varchar | Not null |
| playoffs | Varchar | Not null |

* + - 1. Teams

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Constraint |
| team\_id | Int | Primary key |
| team\_name | Varchar | Not null |
| team\_sym | Char | Not null |

* + - 1. Scores

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Constraint |
| score\_id | Int | Primary key |
| game\_id | Int | Foreign key |
| team\_sym\_away | Char | Not null |
| team\_sym\_home | Char | Not null |
| score\_away | Int | Not null |
| score\_home | Int | Not null |
| Q1\_away | Int | Not null |
| Q2\_away | Int | Not null |
| Q3\_away | Int | Not null |
| Q4\_away | Int | Not null |
| OT\_away | Int | Not null |
| Q1\_hoe | Int | Not null |
| Q2\_home | Int | Not null |
| Q3\_home | Int | Not null |
| Q4\_home | Int | Not null |
| OT\_home | Int | Not null |

* + - 1. Betting Lines

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Constraint |
| betting\_line\_id | Int | Primary key |
| game\_id | Int | Foreign key |
| whos\_favored | Enum('home', 'away') | Not null |
| spread | Decimal(5, 2) | Not null |
| total | Decimal(5, 2) | Not null |
| moneyline\_away | Int | Not null |
| moneyline\_home | Int | Not null |

* + - 1. Betting Results

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Constraint |
| betting\_results\_id | Int | Primary key |
| game\_id | Int | Foreign key |
| id\_spread | Tinyint | Not null |
| id\_total | Tinyint | Not null |

* + - 1. Second Half Lines

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Constraint |
| H2\_line\_id | Int | Primary key |
| game\_id | Int | Foreign key |
| H2\_spread | Decimal(5, 2) | Not null |
| spread | Decimal(5, 2) | Not null |

1. **Business Question Examples**
2. Which teams had the highest average scoring margin during regular season games compared to playoff games?
3. What percentage of games in a season resulted in the favorite failing to cover the spread?
4. By team, identify percentage of games where the underdog covered the second half spread but not the full-game spread.
5. What is the average point differential games played during the playoffs versus regular-season games?
6. What is the average point differential for games played during the playoffs versus regular-season games?
7. Determine the top three teams with the best performance based on win rates in close games decided by 5 points or fewer.
8. Calculate the percentage of games where the total score matched exactly with the line. Include a breakdown by season.
9. Which team had the most games where they covered the spread as an underdog during a specific season?
10. By season and team, identify how often second half betting results differed from full game betting results.
11. Is there a statistically significant difference in average home scores versus away scores during the regular season vs the playoffs, and where?