***Final Project Draft***

***Introduction:***

I am creating a database model to represent the professional chess world. Using real world data to complete the tables, I plan to use queries to identify any noteworthy anomalies in the data.

I chose this topic for two reasons. One of my friends during my undergraduate program convinced me to begin playing chess online during the COVID-19 pandemic; online chess quickly became one of my favorite hobbies, making chess a natural topic for my project. The second reason is the prevalence of drama surrounding cheating in chess lately. With allegations against players being levied frequently and suspicions of the problem being more pervasive than previously known, delving into the relevant data has never been more important.

My pessimistic plan is to construct a complex ER diagram and then complete the tables with some real-world data when possible. The data is to be normalized as well. There would be a dozen tables, including player biographical information, online and over-the-board results, rating histories, players who are suspected of cheating, and more.

My optimistic plan is to make each table thoroughly filled with entirely real-world data, and then to perform queries to identify outliers. For example, I could find standout players who achieve stronger results online than over the board. Performing significance testing would be one way of accomplishing this. Moreover, I could investigate any other ideas that I encounter during exploratory analysis, such as if younger players fare better online compared to older players.

***Exploring:***

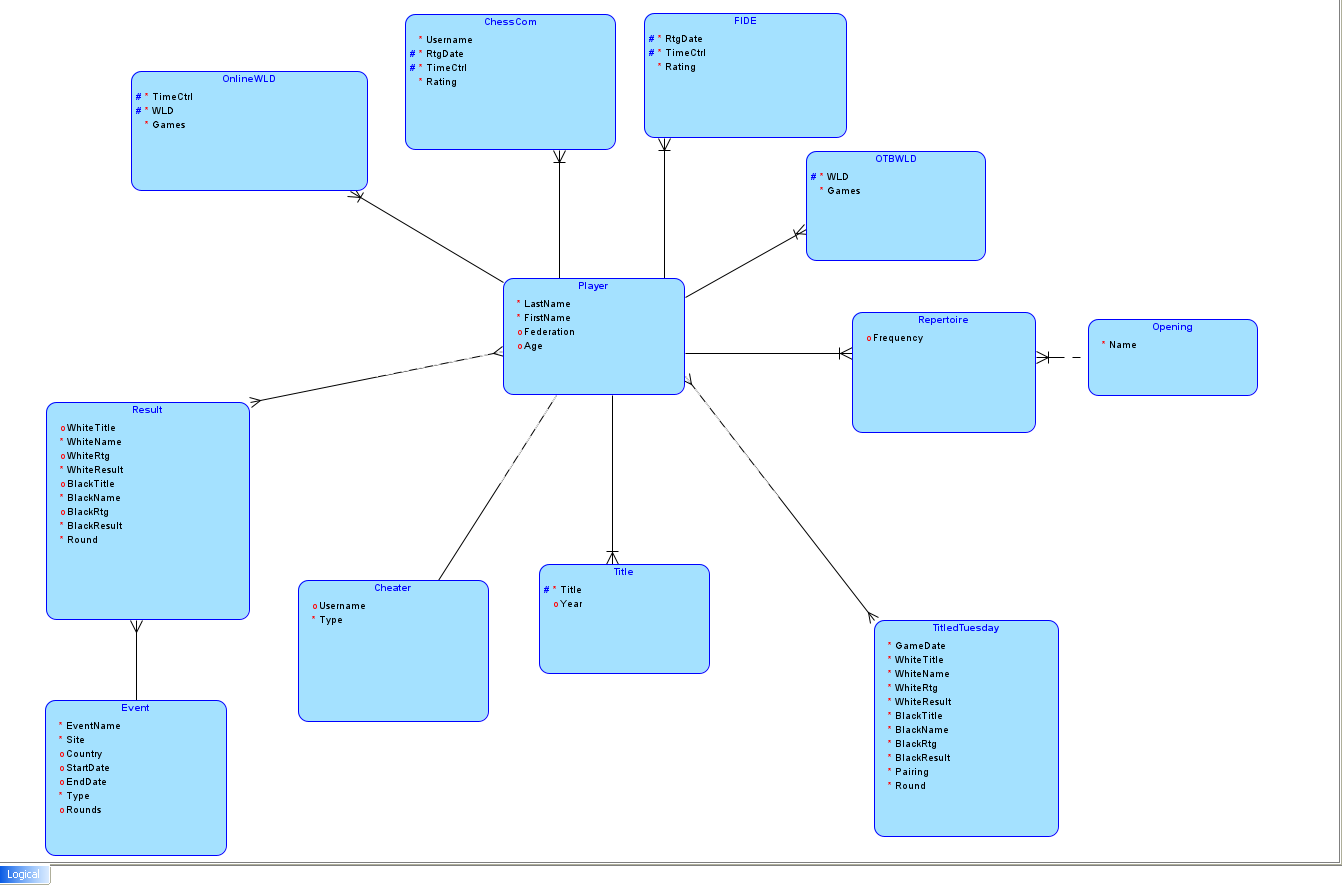
While I had some prior knowledge of SQL, I had very minimal experience working with databases. Therefore, most of my abilities pertaining to this project are ones I have developed during this course. For instance, while I had some familiarity with ER diagrams, I had never made one before, let alone used them to generate actual entities and relationships via DDL (Data Definition Language).

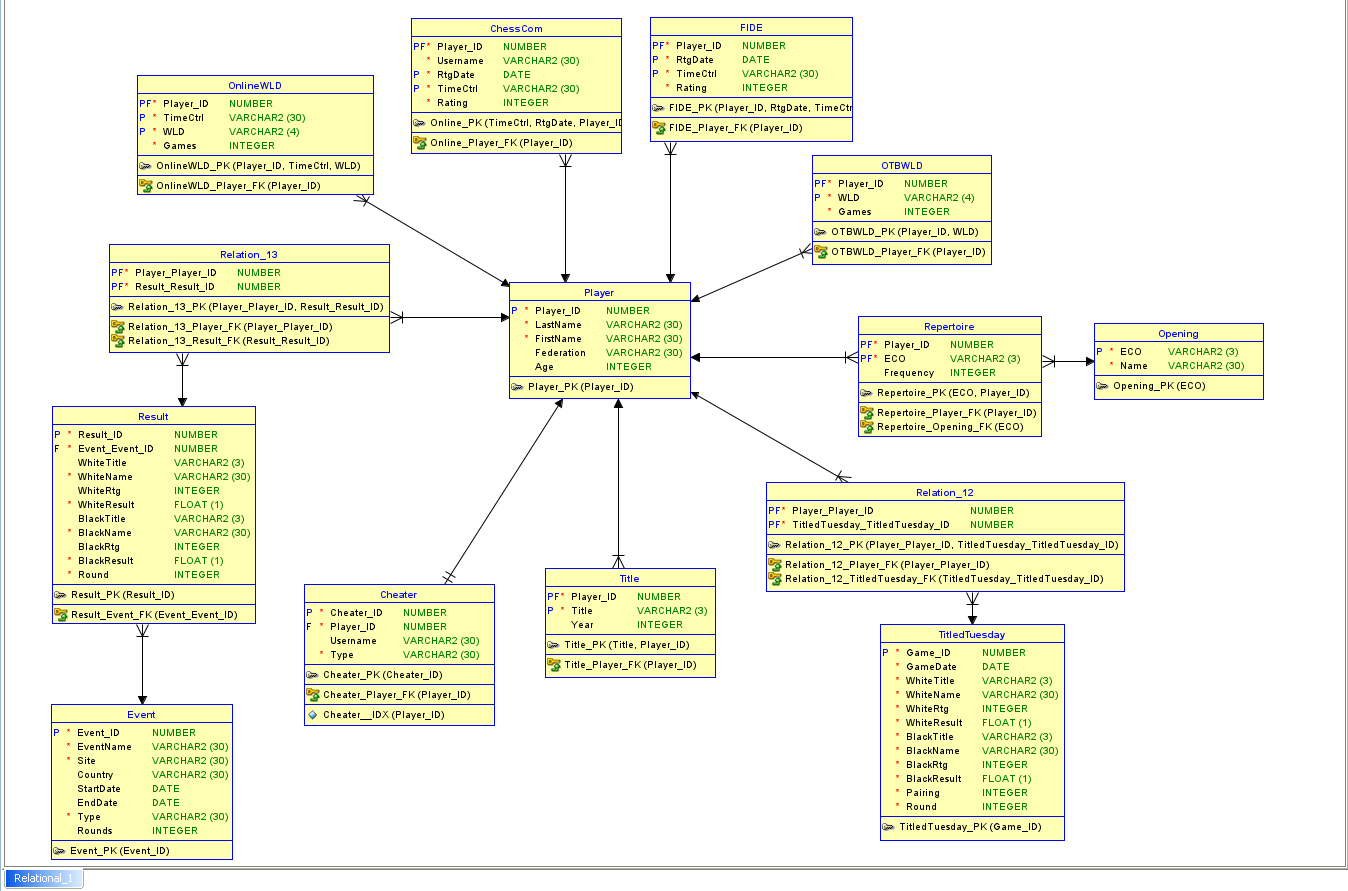
I was always set on making a database model for my final project. The two topics I considered were basketball and chess. However, I settled on the latter due to the abundance of resources in the area that I have recently worked with. My quality understanding of the various sources that I will list below in the following section was the deciding factor.

***Building:***

To build the database, I first considered all the different entities I wanted to include. Of course, I would need to have a player table that would be a foundation for most of the others. I also determined that two separate rating tables, one for over-the-board play and one for online play, would be necessary. Moreover, tables that contain win, draw, and loss rates in each format are also valuable inclusions. A repertoire table contains data on the favored openings of each player, and an opening table gives more information about each opening. An event table contains information on different tournaments, and a result table contains the game results of those events. A Titled Tuesday table contains information specific to the weekly online tournament by the same name, and a title table shows when each player attained their titles. Finally, a cheater table includes players that have been accused of, widely suspected of, and/or admitted to cheating.

From here, I used Data Modeler to create a logical model, which I then forward engineered into a relational model.





My data sources include:

* 2700chess.com for FIDE ratings, over-the-board career statistics, opening preferences, and biographical player information
* Chess.com for online ratings and statistics
* [This dataset](https://www.kaggle.com/datasets/garyongguanjie/chess-com-titled-tuesday-dataset/) from Kaggle that contains Titled Tuesday results from late 2022 into October 2023
* Chess24.com for event information and results

All data was hand-entered, except for the Titled Tuesday results which were imported from a CSV file.

***Discovering:***

One key lesson that I learned was the importance of caution when naming entities and columns. There were several occasions where I was unable to successfully insert data due to the relevant column or table identifiers being invalid. This in turn made dropping the affected tables more difficult.

Another crucial concept is constraints. Constraints are an excellent way to ensure the data entered is all correct, as hand-entered data often contains typographical errors. Adding more constraints to tables via ALTER TABLE statements can make the data insertion process more efficient.