CSC343 Term Project - Phase 3
Data Cleaning Decision

Group members:
Ruijia Wang

The dataset we are choosing for the project is about the teams, players and matches of all seasons of LoL World Champion semi-finals and grand finals. Since the resource of the data is spreaded over the internet, we decided to gather the data and input them into our project by ourselves. Also since the amount of data was not plenty, instead of generating a csv file, we decided to write SQL queries to directly input the data into the schema.

In this case, we started data cleaning at the very beginning of the process. Instead of figuring out NULL values and violated constraints, the data was sorted and edited in order to avoid violations as we were designing the schema and inputting the data. The following examples show how we managed to avoid possible violations:

- **Define proper constraints to each table**: When we were processing the data of LoL teams and leagues, it was noticed that for some teams that established before the existence of any legends, they did not belong to a league. In this case, the “league” values of those teams are NULL values. Therefore, we assigned a foreign constraint of “team” inside the “League” table so that a league could have multiple teams but a team doesn’t have to be assigned to a league.

- **Avoid unnecessary data**: Since we are only working on the teams and players that were the top 4 of each season, other teams and players that never had an experience will be omitted to reduce the workload of data input and also to avoid possible violations.

- **Avoid NULL data by adding values**: In the player table, we noticed that not all birth years of the players are recorded. Therefore, for the players with unknown birth years, we set a value of 0 to the corresponding tuples.