Data Cleaning Steps

1. After the raw data was imported to table “SteamGame” and “SteamData”, we removed all the rows with duplicate names from the raw data table “SteamGame” and “SteamData”. These rows violated some of the constraints of our table so we decided not to include them in this project.
2. For the biggest and most important table “Game”, we removed all the games which did not have NULL value at their ‘developer’ column. We considered these games to have incomplete information because every game should have a developer. We also removed the games which did not have any reviews because they were mostly likely not released yet and therefore, were not related to our research questions.
3. There was not a lot of cleanup to do with tables “Develop”, “Rating”, “Playtime”, and “Language”. We basically just imported corresponding columns into each table.
4. For table “Platform”, we had to check the column “minimum_requirements” of each game to see if they have keywords “Mac”, “Windows”, or “Linux”. If so, we would add the game to the corresponding platform table (temporary table). In the end, we unioned all platform tables and imported them into table “Platform”.
5. For table “Price”, we had to replace strings which are not in the format “$xx.xx” with “$0.00” in the raw data table “SteamGame” (e.g. “Free to Play”). Next, we stripped the “$” sign from each string under the column “original_price” and “discount_price”. In the end, we convert those strings into floating point numbers and imported them along with their “game_id: into table “Price”.