In this data processing task, our primary goal was to clean, fill, and standardize the raw data to ensure data quality and prepare it for further analysis.

First, I examined the dataset's dimensions, which initially contained 13,464 rows and 53 columns. Then, I created a new column, Hours_streams, by summing the Hours and streams columns. After that, I removed rows where Hours_streams was greater than 115 or equal to 0, as they might represent outliers.

For handling missing values, I filled the missing values in the GADE column with the second most frequent category ("Somewhat difficult") and checked the distribution of categories before and after the treatment. For missing values in the streams and Hours columns, I used the mean to fill them. Afterward, I deleted the Hours_streams column since it was only used to filter outliers and was no longer needed.

Next, I cleaned and standardized the League column. First, I converted all values to lowercase and removed leading and trailing spaces. I used regular expressions to extract the ranking information from the values. I then standardized common spelling errors and synonyms (such as "g", "silverii", "platinium") into consistent rank categories. The final valid ranks included: unranked, gold, diamond, bronze, silver, and unspecified. Less common ranks were classified as unspecified.

For other missing values, I replaced 'Unknown' in the Residence column with the most frequent value and filled missing values in the Reference column with 'Other'. For the SPIN1 to SPIN17 columns, I filled missing values with the most frequent value (mode). Additionally, I removed some irrelevant columns, such as Residence and accept.

For text columns (such as Playstyle, earnings, and whyplay), I cleaned the text by removing punctuation, converting to lowercase, and stripping spaces. I then grouped less frequent values into the category Other, replaced Other with NaN, and deleted those rows.

When handling the earnings and whyplay columns, I merged low-frequency values into main categories. For example, earnings was classified into: "playing for fun," "playing for fun but earning a little on the side," and "earning a living by playing." Similarly, whyplay was divided into: "fun-related" (such as having fun, relaxing) and "goal-related" (such as improving, winning).