Data Cleaning Steps in Excel

1. Dropped the following columns:
    - LEAGUE_NAME
    - LEAGUE
    - FINISHED

2. Changed values in SEASON, e.g. “2005” to “2005/2006”. There are always two different years to a season. The year itself can be derived from the MATCH_DATE column.

3. Filled all blank cells in column VIEWER with 0.

4. Replace two cities in LOCATION column due to ambiguity:
    - Gladbach → Mönchengladbach
    - Frankfurt → Frankfurt am Main

5. Created a new sheet “Teams” with the following columns:
    - TEAM_ID: Contains all distinct values from HOME_TEAM_ID and AWAY_TEAM_ID
    - TEAM_NAME: Contains all distinct values from HOME_TEAM_NAME and AWAY_TEAM_NAME
    - TEAM: Contains all distinct values from HOME_TEAM and AWAY_TEAM
    - TEAM_LOCATION: Contains the city of each team from LOCATION
    - TEAM_ICON: Contains all distinct values from HOME_ICON and AWAY_ICON

    After this, only the columns HOME_TEAM_ID and AWAY_TEAM_ID remained in the main table, which was renamed to “Matches”.

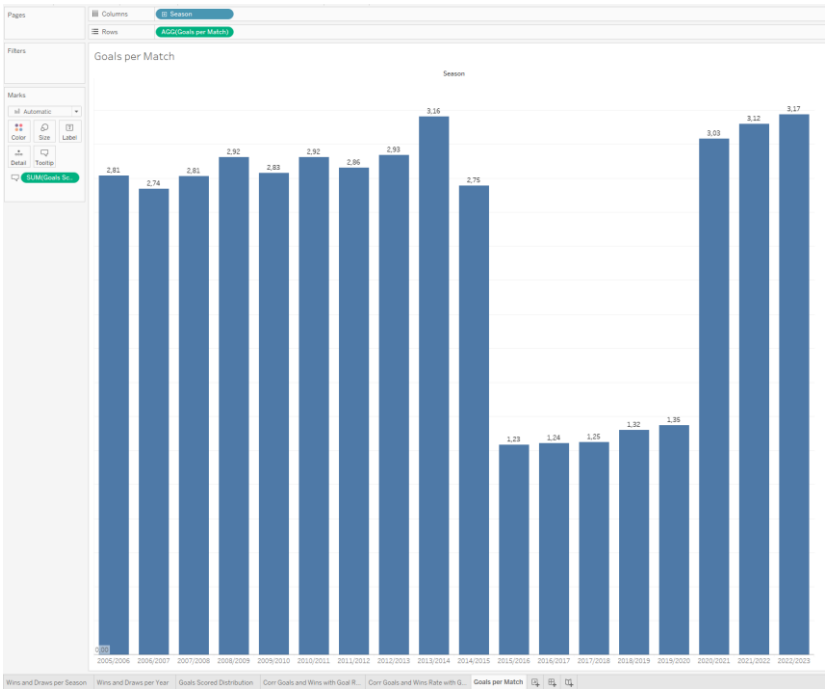
6. Used Power Query’s Unpivot-Feature to split each row into two separate rows, separating information about the home and away team separately. This allows to change the columns GOALS_HOME, GOALS_AWAY, DRAW, WIN_HOME and WIN_AWAY to GOALS_SCORED, GOALS_CONCEDED, WIN, LOSS, DRAW. Example:

MATCH_ID	MATCH_DATE	SEASON	LOCATION	VIEWER	MATCHDAY	MATCHDAY_NR	HOME_TEAM_ID	AWAY_TEAM_ID	GOALS_HOME	GOALS_AWAY	DRAW	WIN_HOME	WIN_AWAY
0	05.08.2005 20:30	2005/2006	München	0	1. Spieltag	1	40	87	3	0	0	1	0

↓

MATCH_ID	MATCH_DATE	SEASON	LOCATION	VIEWER	MATCHDAY	MATCHDAY_NR	ROLE	TEAM_ID	GOALS_SCORED	GOALS_CONCEDED	WIN	LOSS	DRAW
0	05.08.2005 20:30	2005/2006	München	0	1. Spieltag	1	HOME	40	3	0	1	0	0
0	05.08.2005 20:30	2005/2006	München	0	1. Spieltag	1	AWAY	87	0	3	0	1	0

7. During exploratory data analysis it was found that the data of seasons 2015/2016 to 2019/2020 must be in error because the total number of goals differed significantly from the years before and after:



For that reason new data were scraped for the affected period from https://www.fussballdaten.de/bundesliga (see Jupyter Notebook “Webscraping Bundesliga.ipynb”). The new data was used to overwrite the wrong data in the existing file to fix the issues: