UDACITY

Investigation of a Dataset

Project Work – Soccer Database

Jacek Niedzielski

Vienna, May 2020

Table of Contents

Table of Figures

1. Short description of dataset and stated questions

The dataset that I have chosen for my investigation is the *"Soccer Database"*.

The dataset is available on *kaggle* an consists of many tables like leagues, matches, teams and their attributes and so on.

After a short inspection of tables right on the website, I have decided to analyze the teams and their progresses among the seasons, as well as odds offered by different bookmakers. These aspects are not related to each other in my analysis, hence I have prepared separate reports for them.

- Betting.ipynb for analysis of bookmakers
- Teams_attributes_influence.ipynb for analysis of football teams

In my analyses I have tried to answer the questions:

- which bookmaker has the best rate of return
- which teams did the best progress in the last seasons and which of their attributes decided whether they have been successful or not.

2. Used tools

I have used following tools for project completion:

- DB Browser for first insights of data
- Python 3.7 within JupyterNotebook environment for data analysis
- Microsoft Word 2010 for reporting

3. Short insight and SQL queries

The very first look at the data can give a bit of overview:

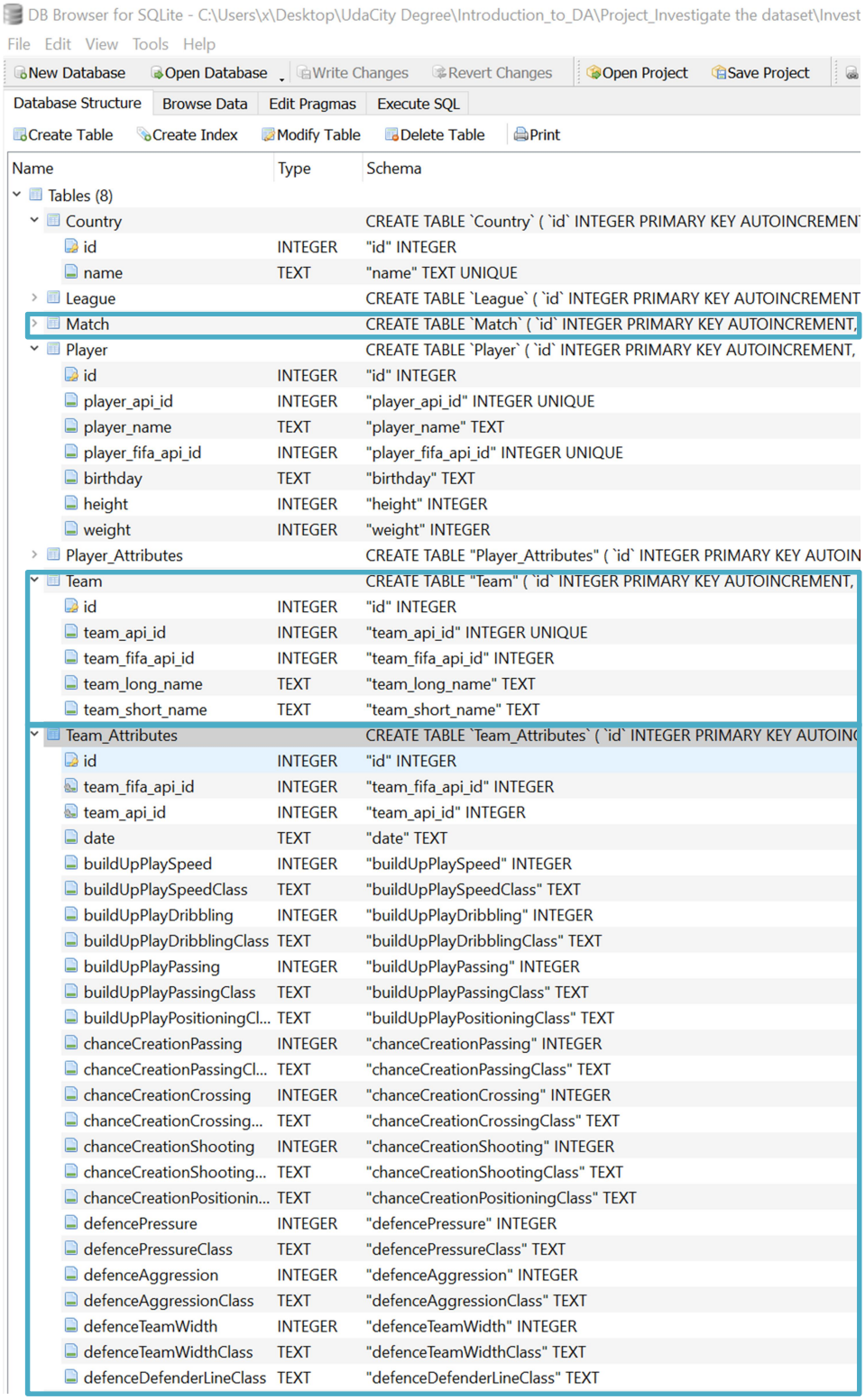

Fig. 1: Overview of the "European Soccer Database" in DB Browser

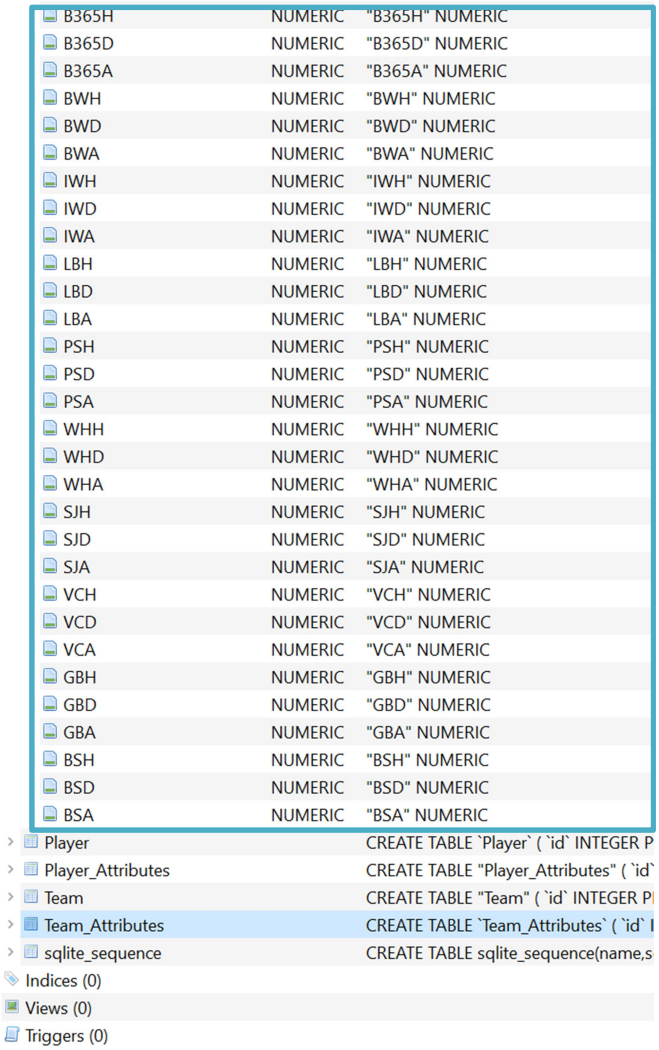

Fig. 2: Last columns of "Match" table

In order to gain the required csv files I used the following queries:

```
SELECT *
FROM Match
```

and:

```
SELECT *
FROM Team
JOIN Team_Attributes
ON Team.team_api_id = Team_Attributes.team_api_id
```

I have saved the information in the following files, which I have included in the zip data.

Matches.csv , Teams_with_Attributes.csv

4. Further steps

The whole analysis including more specific description of questions, simplifications, conclusions and whole data exploration and cleaning can be found in the notebooks being part of the zip file