Jean-Claude du Sart
Career Foundry
Data Analytics Immersion

Achievement 6

Exercise 6.1: Sourcing Open Data

Directions

1. If you haven't done so already, download your Achievement 6 project brief (.pdf)
2. The data you use for your project will need to meet certain criteria as defined in the brief. Read through the data requirements now to be sure the data you choose is appropriate.

- Be open source.

  The Data I have chosen I got from Kaggle. Kaggle is a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC. The data is Public Domain

- Come from an authentic/authoritative source.

  The data is gathered from several sources including but not limited to Wikipedia, rsssf.com, and individual football associations' websites.

- Include non-anonymized column names.

  The data does not have any personally identifiable information (PII). It has just statistics.

- Be no more than 3 years old (up to a maximum of 10 years if you've found a perfect data set for your needs and no newer data is available).

  The data ranges from 1872 up until 2024, but we can use an extract of it for the last 10 years and just compare to that of the history.

- Contain at least 2 continuous variables (excluding index or ID variables, dates, years, etc.).

  "home_score" and "away_score" are 2 variables that can hold the amount of goals for each team and can take on any non-negative inter value.

- Contain at least 2 categorical variables (excluding index or ID variables, dates, years, etc.).

  tournament and country are both considered to be categorical variables.

- Contain at least 1,500 rows.

  Main database contains 49019 records.

- Include a geographical component with at least 2 different values (e.g., countries, continents, U.S. states, cities, latitude, and longitude values—anything you can visualize on a map!).

  countries and neutral (True/False) ground played on.

3. Source your data. Use the requirements (and your own interests) to source an open data set from the web. We introduced you to several sources in this Exercise but feel free to look elsewhere, as well.

   https://www.kaggle.com/datasets/patateriedata/all-international-football-results - **International football results from 1872 to today**

   https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017 - **International football results from 1872 to 2023**

4. Document created.

5. Create a "Data Source" section in your project document and provide the following information:

   - A summary of your data source. We recommend you revisit Exercise 1.4: Sourcing the Right Data for a recap on what to include in your summary.
   - An explanation for why you've chosen this data set.

Data Source

This dataset includes **49019** results of international football matches starting from the very first official match in 1872 up to 2023. The matches range from FIFA World Cup to FIFI Wild Cup to regular friendly matches. The matches are strictly men's full internationals, and the data does not include Olympic Games or matches where at least one of the teams was the nation's B-team, U-23 or a league select team.

`all_matches.csv`

date: The date when the match took place.
home_team: The team playing at their home venue.
away_team: The team playing as the visitor.
home_score: The number of goals scored by the home team.
away_score: The number of goals scored by the away team.
tournament: The name or type of the tournament or competition.
country: The country where the match took place.
neutral: Indicates whether the match was played on neutral ground (Boolean: True or False).

The below are additional columns I have added for further Analaysis
year: The year in which the match occurred.
total_goals: The total number of goals scored in the match (sum of home and away scores).
result: The overall result of the match (win, lose, or draw).
home_result: The result for the home team (win, lose, or draw).
away_result: The result for the away team (win, lose, or draw).

`countries_names.csv`

original_name - country name used in the dataset.
current_name - country name used nowadays (or name of the country that inherited the history).

I have chosen this dataset as a result of being a football enthusiast and thought it would be fun to gather insights into matches and history of the game over time and also to see how different continental countries performed.

6. Clean your data. Conduct some basic data cleaning and consistency checks in Jupyter to ensure your data is ready for further analysis.

   Data Cleaning and Consistency Checks.ipynb (Will be uploaded with this document)

7. Understand your data. Develop a basic understanding of your data set by reviewing the variables and performing basic descriptive statistical analysis. You might want to make a data profile similar to what you did in Achievement 1.

Data Profile

Foot Ball Results dataframe

Variables	Description	time-variant/ invariant	structured/ unstructured	qualitative/ quantitative	qualitative: nominal/ordinal quantitative: discrete/continuous
date	The day, month and year of the match	invariant	structured	quantitative	discrete
year	The year the match was played	invariant	structured	quantitative	discrete
home_team	The name of the home team	invariant	structured	qualitative	nominal
away_team	The name of the away team	invariant	structured	qualitative	nominal
home_score	the score of home team as an integer	invariant	structured	quantitative	ordinal and discrete
away_score	the score of away team as an integer	invariant	structured	quantitative	ordinal and discrete
total_goals	the total goals added by both teams as an integer	invariant	structured	quantitative	ordinal and discrete
result	String of home team, "draw"or away team	invariant	structured	quantitative	ordinal and discrete
home_result	home result as "WIN", "DRAW" OR "LOSE"	invariant	structured	quantitative	ordinal and discrete
away_result	away result as "WIN", "DRAW" OR "LOSE"	invariant	structured	quantitative	ordinal and discrete
tournament	type of match	invariant	structured	qualitative	categorial
country	Location of match	invariant	structured	qualitative	categorial
neutral	Is the ground at home team location	invariant	structured	qualitative	categorial

8. Consider limitations and ethics. Outline any limitations and ethical considerations presented by the content of your data, its source, and/or how it was collected.

- The data is limited to real time situations and need to be manually updated which could lead to errors in capturing. I doubt any bias could be relevant in the results captured as this is from real life statistics.
- There are possible missing data points in the goal scorer's spreadsheet as I picked up when doing some cleaning that some of the names were missing, this was around 200 in a database of 40000+ records.
- The data is collected frequently and checked constantly. It has been collected/gathered from several sources including but not limited to Wikipedia, rsssf.com, and individual football associations' websites.
- The data is for statistical purposes only and does not have any ethical considerations to be weary of.

9. Include the results of steps 6 to 8 in a second section of your project document. This second section can be titled something like "Data Profile."

Completed above.

10. Define questions to explore. In a third section of your project document, define a list of questions to explore with your analysis. As mentioned in the Exercise, you may want to revisit Exercise 1.2: Starting with Requirements for a recap on writing good questions.

- Who is the best team of all time? Have individual goal scorers contributed to this?
- Which teams dominated different eras of football?
- What trends have there been in international football throughout the ages - home advantage, total goals scored, total goals conceded.
- Can we say anything about geopolitics from football fixtures?
- Which countries host the most matches where they themselves are not participating in?
- How much, if at all, does hosting a major tournament help a country's chances in the tournament?