CS5200 Group Project 3 Appendix

Group 3 – Calvin Lo, Hao Niu, Tianyu Fang, Shitai Zhao (Stanley)

This appendix section provides additional information that is relevant to the project, with an aim to offer to any potential readers who may be interested in the project, an overall better understanding as to how the project was initiated, designed, and executed.

Introduction

This project is a data analysis project that aims to explore the performance of the host country in the Olympic Games, from Athens 1896 to Beijing 2022. The project was initiated by a group of students from the CS5200 Database Management Systems course at Northeastern University, as part of the course project in the summer semester of 2024. A dataset containing information about the Olympic Games, including the host countries, medal counts, athletes, and results information was retrieved from Kaggle, and was used as the primary data source for the project.

The current project is structured to include the following sections:

- Conceptual & Logical modeling
- Data realization & implementation
- Data cleaning & loading
- Data querying & trend analysis

Remaining sections of the appendix will provide further details on rationales behind topic selection, project focal point, Data selection, and Data cleaning process.

Topic Selection

Despite the fact that Paris 2024 is taking place at the moment just 8000 KM away, the topic of results & medals analysis of the Olympic Games (host countries) was not an immediately go-to choice for the team as the project started.

We firstly considered other topics mostly related to the health sector, but later found that the datasets explored were not as representative and engaging as we would like the project to be. Also, after a few discussions with the course instructor, we realized as well that the heart diseases dataset which we initially considered may not contain the information comprehensive enough to either justify our hypotheses, or to support the possible conclusions that we may draw from the data.

After some rounds of selections, we eventually landed on the Olympic Games dataset from Kaggle, which is at least in our opinion, a worthwhile topic to explore and likely more relevant to not only us but also the general audience.

Focal Point: Host Country (Effect)

The following hypothesis were formulated in the project:

- Countries win more medals when hosting the Olympics.

We understand the implications the word "Olympic Games" may suggest to the general audience, and choosing to primarily focus on the host country's performance in the Olympic Games was a strategic decision made based on the availability of the data, the potential interests of the reader, and mostly importantly the practical considerations such as the time constraints and the complexity, inconsistency of the original dataset.

The idea of "Host Country Effect", despite from time to time being mentioned in the project, was not originally proposed by the team, but rather a concept discovered during the research phase. We acknowledge the intrinsic complexities and the potential biases that may be associated with this term, from definition to measurement, but it is not the intention of the current project to delve into the possible philosophical and social debates surrounding it, but rather to provide a simple, yet informative analysis of the host country's performance in the Olympic Games. And for most of times when we refer to the "Host Country Effect", it can be understood as a general term which represents the differences in medal counts across different host countries and non-host countries.

Data Selection

As reflected from the original datasets and schema design diagrams, the data selection process was not as straightforward as it seems to be. The original dataset, despite being comprehensive and detailed, suffered from continuous inconsistencies, missing values, and other data quality issues, and all the factors lead to the final data schema being a subset of the original, excluding certain data related to athletes, results, and events, not only because such information was deemed to be less relevant to the project, but also because the data quality issues were too severe to be resolved within the time constraints of the project(Please see examples from Data Cleaning section in the notebook).

But overall, the data selection process was based on the following criteria:

- Relevance: The data selected should be relevant to the project, and should be able to support the hypotheses formulated in the project. I.e., are the data selected relevant to the project's focal point previously discussed?

- Completeness: The data selected should be complete, and should not contain missing values or other data quality issues that may affect the analysis. I.e. will the data used to be enough to find the trends and patterns in medal counts differences?

- Consistency: The data selected should be consistent, and should not contain inconsistencies or other data quality issues that may affect the analysis. I.e., are the data selected consistent in terms of format, structure, and content (country name, event title etc.)?

- Availability: The data selected should be available, and should be accessible to the project team. I.e., are the data selected available in the original dataset, or do they need to be retrieved from other sources?

Data Cleaning

The data cleaning process was arguably the most challenging part of the project, as previously mentioned, the original dataset contained a large amount of data quality issues, including missing values, inconsistencies, and other errors that needed to be resolved before the data could be used for analysis, and in our case specially, event titles and country names were the most problematic fields. A significant amount of effort was put into manually mapping the attributes to the correct values, please refer to section 3 of the notebook for more details.

Overall, a series of data cleaning steps were performed on the original dataset, which trimmed down the original from 4 CSV files, over 34 columns and 300000 rows, to a more manageable size which involved 3 CSV files, 14 attributes, and around 150000 rows. All above mentioned data selection and cleaning criteria were taken into consideration to ensure the finalised dataset was consistent, relevant, and sufficient to answer the hypotheses proposed, and for detailed results analysis, please refer to section 6 from the notebook.

Conclusion

To conclude, the project on the Host Country Analysis of Olympic Games was designed, implemented and analyzed to provide a series of insights of mostly medal counts of the host countries in the Olympic Games, from Athens 1896 to Beijing 2022. There are certainly limitations and constraints associated with the project, such as data inconsistencies, time constraints, and the complexity of the topic itself, but the project team hopes that the project will be able to provide some valuable information and insights to the general audience, and to inspire further research and analysis in the field of sports analytics and data science.

References

[1] Petro. “Olympic Summer & Winter Games, 1896-2022.” *Kaggle*, 11 Apr. 2022, www.kaggle.com/datasets/piterfm/olympic-games-medals-19862018/data.