Data Warehousing

Olympic Medals Datawarehouse Project

Chensu Yang

From

Master of Information Technology (CITS 5504)

14 April 2024

Contents

1 Introduction and Design

1.1 Data Information

This project is supported by 7 CSV files encompassing various subjects such as the economy, population, life expectancy, countries and regions, mental health, Olympic hosts, and Olympic medals. 6 of these files will be utilized for data warehousing and association rule mining purposes. To meet client demands regarding economic data consultation, an additional external data file has been incorporated into this process, identifiable within a zip file prefixed with "API".

1.2 Data Warehouse Design

1.2.1 The reason Economic Data.csv was abandoned

The decision to exclude "Economic Data.csv" was made for the following reasons:

- The file is focus on data from the year 2020 significantly limits its utility. Our data warehouse design emphasizes time-series analysis, and the inclusion of data from only one year fails to accurately represent the economic status of countries during each Olympic Games, excluding that of 2020.
- World GDP US dollar, A time based GDP data was found for the extension of economic data satisfy the need from clients and the data warehouse design.

1.2.2 Brief for dimension design

9 dimensions are created for designing the data warehouse, including:

1. "dimcountry" This dimension encapsulates geographical data, including country names and regions, and provides an alternative key to denote country codes.
2. "dimeventattendtype" Includes information on the gender of participants and the type of participation (individual or team).
3. "dimeventcategory" Comprises details on the disciplines and the specific categories of Olympic events.

4. "dimgamehost" Details the host city, season, and the names of each Olympic event.

5. "dimgdp", "dimlifeexpectancy", "dimmentalillness", "dimpopulation" These dimensions are responsible for holding data on various indices, such as GDP percentages, life expectancy, mental health statistics, and population figures, relevant to the countries of the athletes participating in the Olympics.

6. "dimtime" Captures year data, segmenting it into decades to facilitate time-series analysis and trend observation within the warehouse.

1.2.3 Business Queries

The clients might ask for queries from their business needs in order to improve their Olympic performance or other investment needs. The virtual clients could be listed as below:

1. The Australian Olympic Committee (AOC) is a non-government, not-for-profit organisation, committed to the development of youth and sport. It is our responsibility to select, send and fund Australian Teams to the Olympic Games. AOC might aim to find the information focus on Australian athletes' historical data to improve Australian team's strength.

2. The International Olympic Committee (IOC) is a non-governmental sports organisation based in Lausanne, Switzerland. IOC encourage and support the promotion of women in sport in pursuit of equality between men and women and encourage and support measures relating to the medical care and health of athletes. IOC might aim to explore the world wide data of Olympic performance.

3. The European Union (EU) is a supranational political and economic union of 27 member states that are located primarily in Europe. EU may have the requirement of making investments on collecting data for Olympic performance to make the decision of the fundings of European Olympic Committees and may ask for some requirement for the adjustments of European Games to increase their power in Olympics.

The different business queries may come from those clients. 5 queries would be received from every client as below:

1. AOC: In which ten disciplines do Australian athletes consistently achieve the highest performance?
2. AOC: Can we observe a trend in Australia's overall historical performance in the Olympic Games or winter games by years?
3. IOC: Could you provide a visualization that represents the distribution of genders across all Olympic events by years?(Not including Open and Mixed)
4. IOC: Is there a correlation between the mental well-being of athletes and the acquisition of gold medals?
5. EU: What are the ten disciplines where European athletes tend to have the least success?
6. EU: How economy impress the performance of Olympics?
7. AOC: How many gold medals has Australia won in each decade of the Olympic Games?
8. AOC: How do Australia's Olympic medal counts compare to those of New Zealand across all Olympic Games?
9. AOC: What is the contribution of Oceania countries in terms of Olympic medal?
10. IOC: Is there a correlation between a country's population size and its number of Olympic gold medals?
11. IOC: Is there a correlation between a country's life expectancy and its number of Olympic gold medals?
12. IOC: How do the continents compare in terms of their Olympic medal contributions by decades?
13. EU: What are the gold medal statistics for EU countries in the Olympic Games?

14. EU: How have EU gold medal wins varied by decade in the Olympic Games?
15. EU: How does the gender distribution of Olympic medals compare between North America and Europe?

Question: The reasons why the identified client(s) are important?

- AOC, IOC, and the EU are stakeholders in the Olympics, each with distinct perspectives and concerns. The AOC, as the participant place, focuses on the performance and strengths of its athletes. The IOC, as the organizing authority, is interested in the broader scope of the games, including participant demographics and the welfare of athletes. Meanwhile, the EU, as an investor, analyzes performance from a financial and regional success standpoint. The queries designed for the reflections of their unique interests, showing a well-considered approach to addressing their potential needs in understanding different facets of the Olympic Games.

1.2.4 StarNet Model

After designing the dimensions and attributes and added them into hierarchies, a StarNet diagram could be built for the further analysis. For answering the business queries above, StarNet footprints should includes essential attributes.

1.2.4.1 Query 1 The query is: In which ten disciplines do Australian athletes consistently achieve the highest performance?

Client's intention is to understand their strengths in the Olympics, possibly to increase their investment in these strengths, and to identify some overlooked strong areas to arrange subsequent financial support or increase the educational support.

The StarNet footprint should includes Australia by name in country dimension and to find 10 top disciplines by discipline in the dimension. (See Fig 2)

1.2.4.2 Query 2 The query is: Can we observe a trend in Australia's overall historical performance in the Olympic Games or winter games by years?

Client's intention could be deciding whether to increase the efforts on Olympics

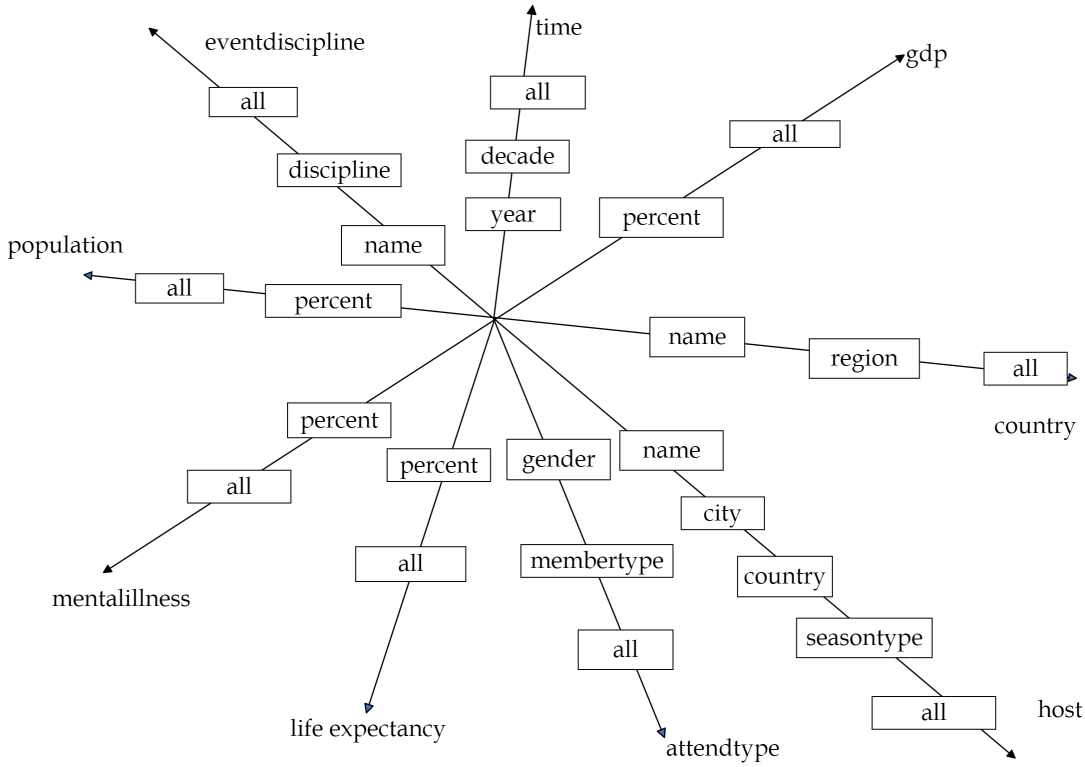


Figure 1: StarNet Diagram

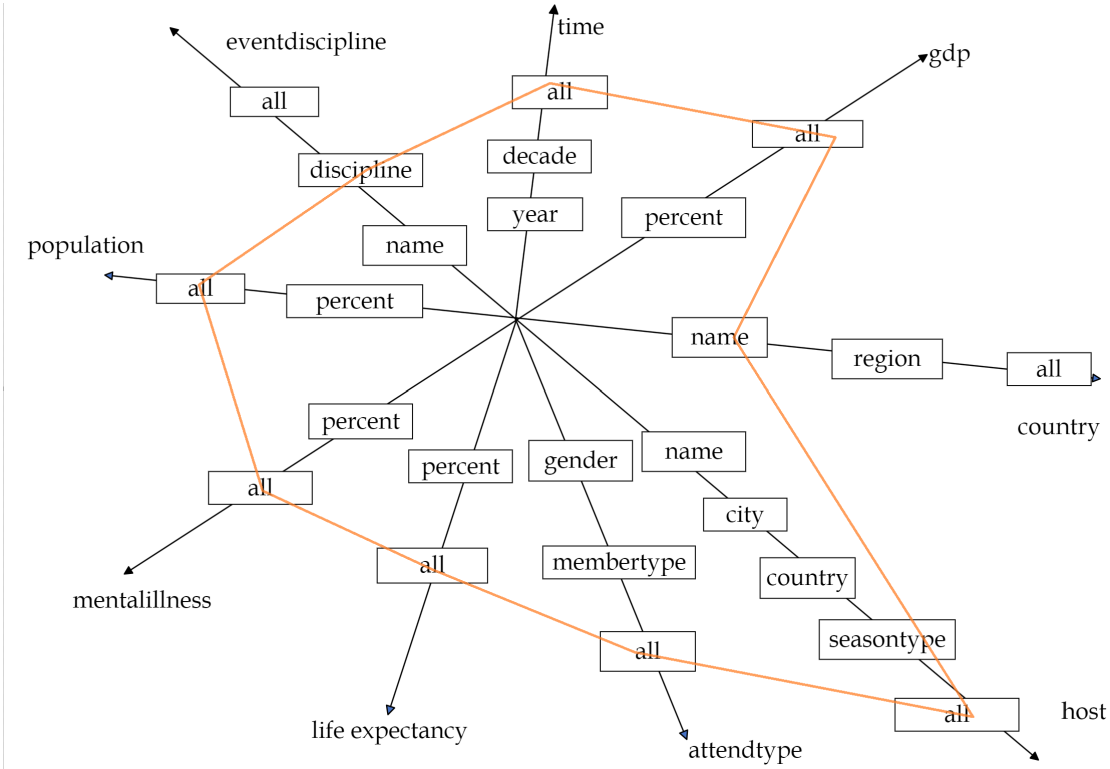


Figure 2: Query 1

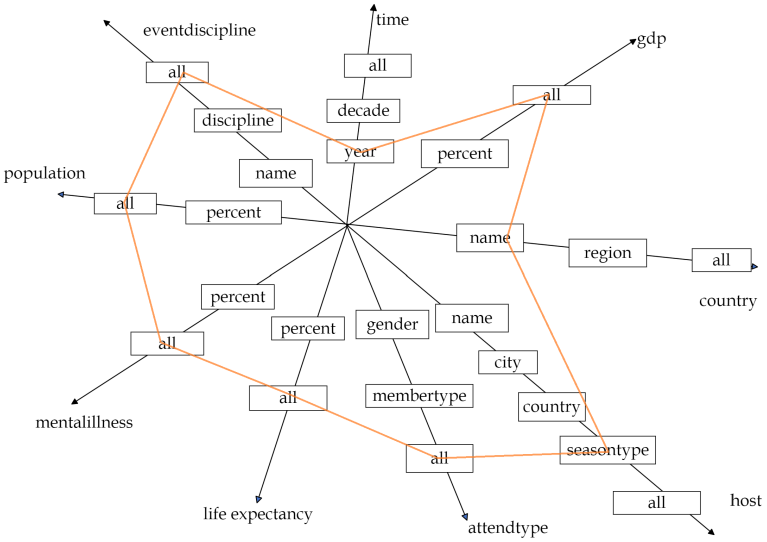


Figure 3: Query 2

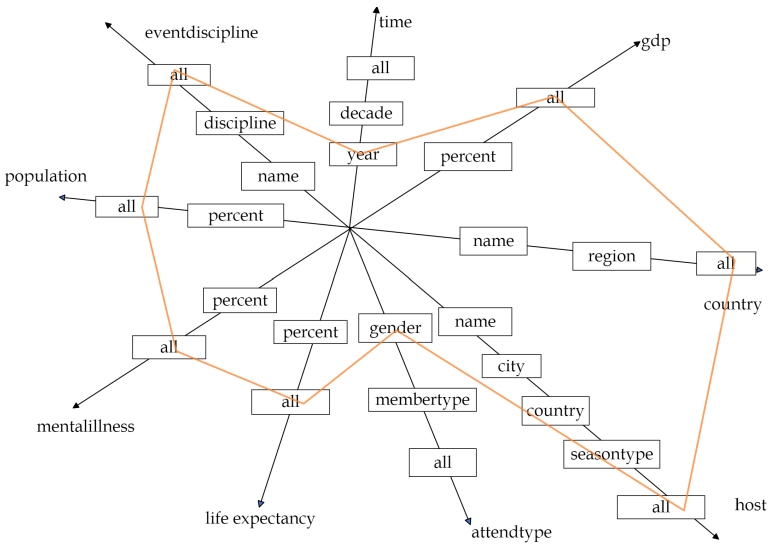


Figure 4: Query 3

based on the historical trend(Financial etc..), like if the trend is heading down increase the financial support but if it increased then maintain the current level.

The StarNet footprint should includes Australia by name in country dimension, summer or winter game by seasontype and seeking historical yearly data from year. (See Fig 3)

1.2.4.3 Query 3 The query is: Could you provide a visualization that represents the distribution of genders across all Olympic events by years?(Not including Open and Mixed)

Client is looking to analyze the historical yearly performance in the Olympics to identify the current situation of gender equality in Olympics.

The StarNet footprint should includes gender in attendtype dimension and seeking historical yearly data from year. (See Fig 4)

1.2.4.4 Query 4 The query is: Is there a correlation between the mental well-being of athletes and the acquisition of gold medals?

Client may want to know how mental wellness related to the performance in

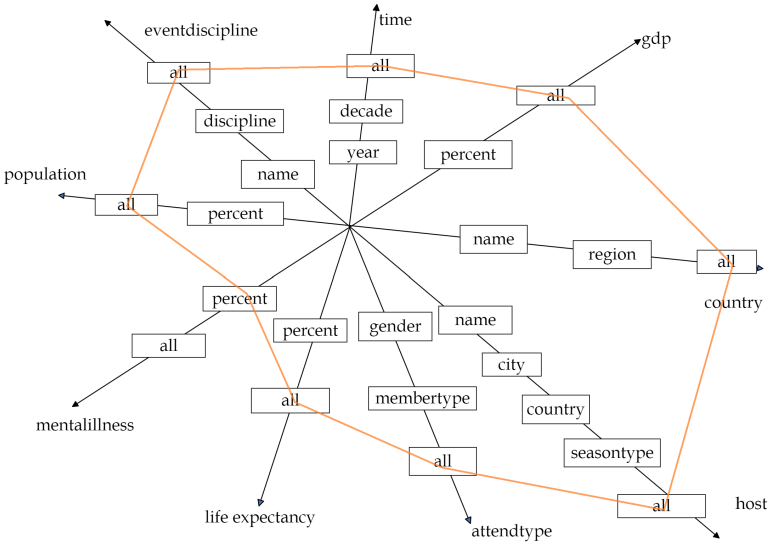


Figure 5: Query 4

Olympics and understanding the logic behind the data.

The StarNet footprint should includes percent from mental dimension to check the mental illness status of the people worldwide. (See Fig 5)

1.2.4.5 Query 5 The query is: What are the ten disciplines where European athletes tend to have the least success?

Client could be interested in the weakness of European athletes and probably to set more events in the European regional games or more financial investment in the European Olympic Committee.

The StarNet footprint should includes discipline in eventdiscipline to find discipline information and region in country to find Europe. (See Fig 6)

1.2.4.6 Query 6 The query is: How economy impress the performance of Olympics?

Client may want to know if the investment would make a progress in the Olympics refer to the gdp influence.

The StarNet footprint should includes percent in gdp to know the gdp status of

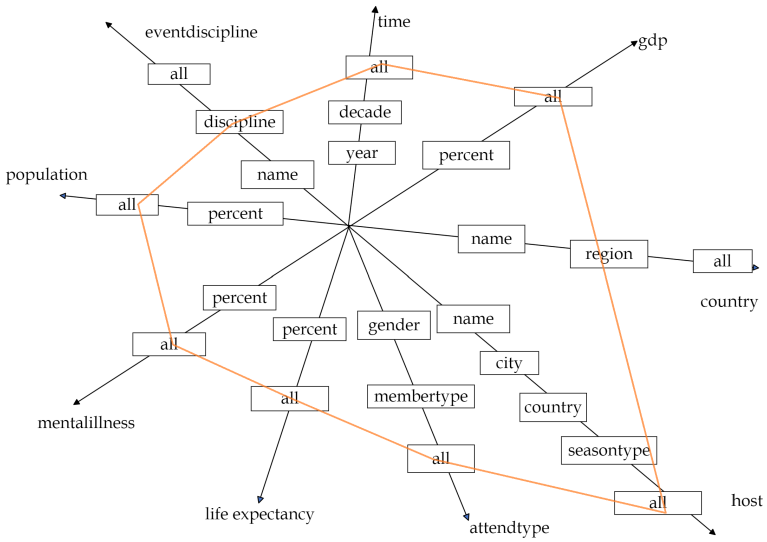


Figure 6: Query 5

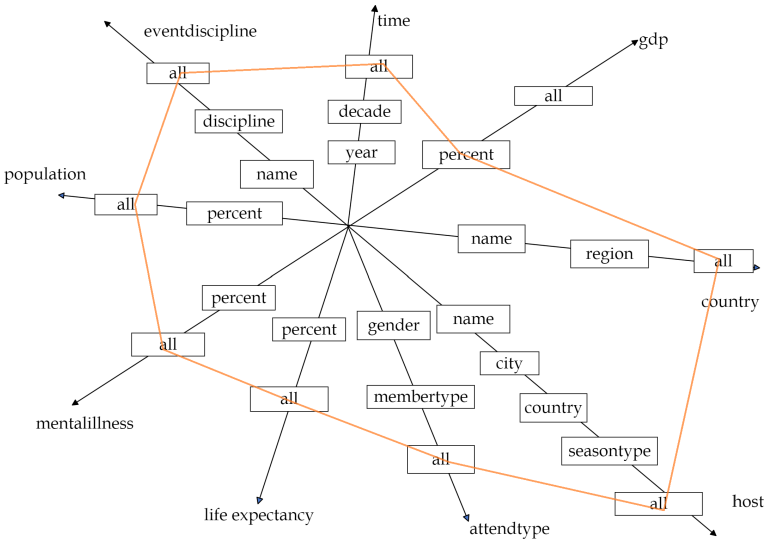


Figure 7: Query 6

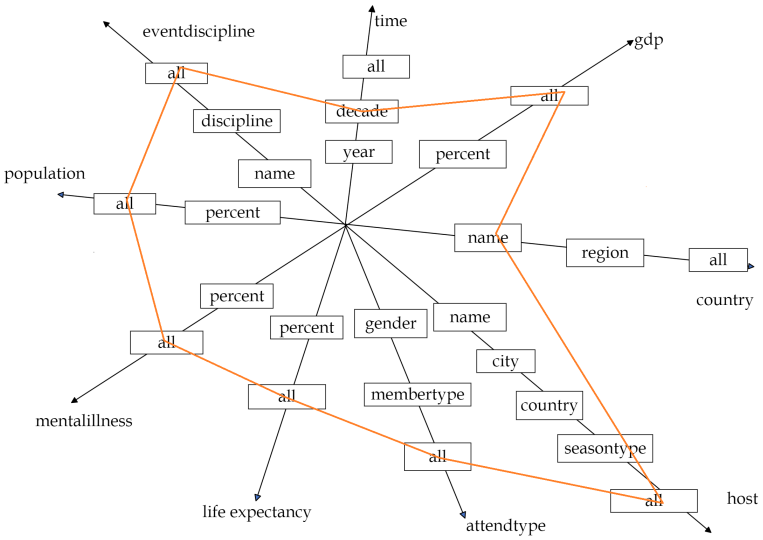


Figure 8: Query 7

each game. (See Fig 7)

1.2.4.7 Query 7 The query is: How many gold medals has Australia won in each decade of the Olympic Games?

The AOC may want to analyze trends in Australia's Olympic performance over the years to inform strategies for athlete development and resource allocation.

The StarNet footprint should includes decade and name of Australia as country name. (See Fig 8)

1.2.4.8 Query 8 The query is: How do Australia's Olympic medal counts in genders compare to those of New Zealand across all Olympic Games?

This comparison could help the AOC understand their regional gender equality of their sports programs relative to a close neighbor country.

The StarNet footprint should includes gender and country name. (See Fig 9)

1.2.4.9 Query 9 The query is: What is the contribution of Oceania countries in terms of Olympic medal?

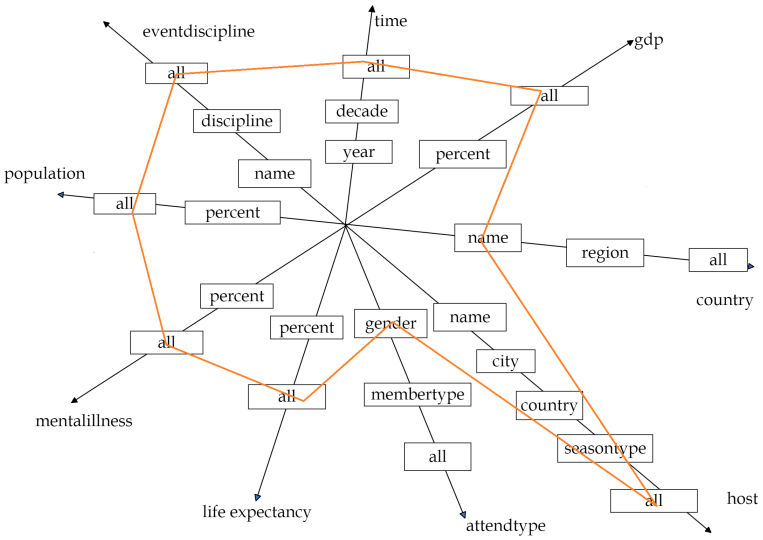


Figure 9: Query 8

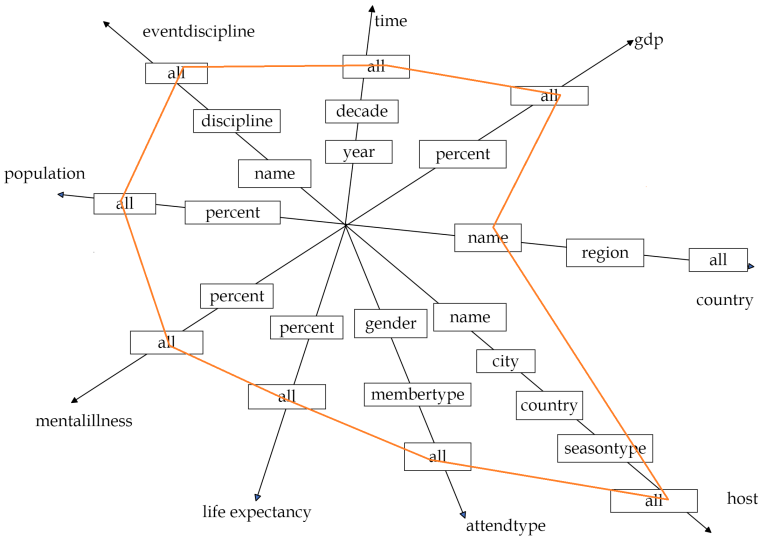


Figure 10: Query 9

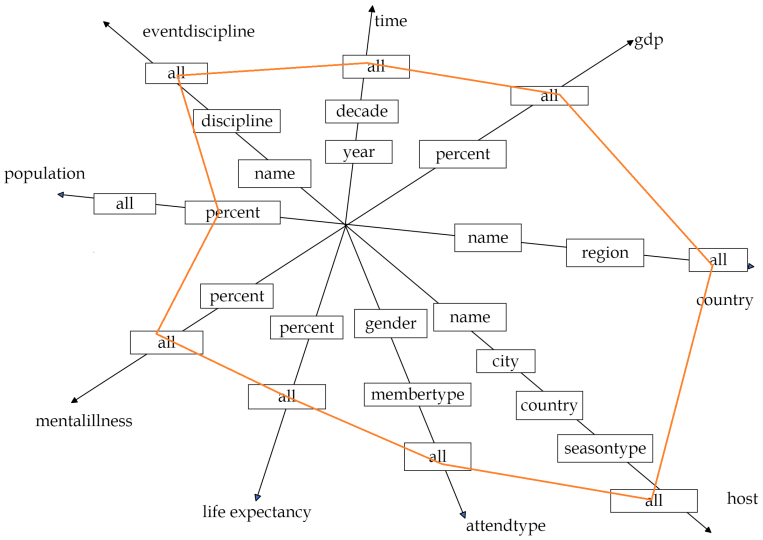


Figure 11: Query 10

Knowing the overall contribution of Oceania countries can help the AOC itself within regional collaborations and competitive strategies, such as exchanging members.

The StarNet footprint should includes country name with filtering Oceania countries. (See Fig 10)

1.2.4.10 Query 10 The query is: Is there a correlation between a country's population size and its number of Olympic gold medals?

The IOC might explore this relationship to assess how population influence national sports success, which could help for redesign the rules.

The StarNet footprint should includes population percent. (See Fig 11)

1.2.4.11 Query 11 The query is: Is there a correlation between a country's life expectancy and its number of Olympic gold medals?

IOC may want to know if better health expectancy outcomes correlate with sporting success, which may highly related to physical health.

The StarNet footprint should includes life expectancy percent. (See Fig 12)

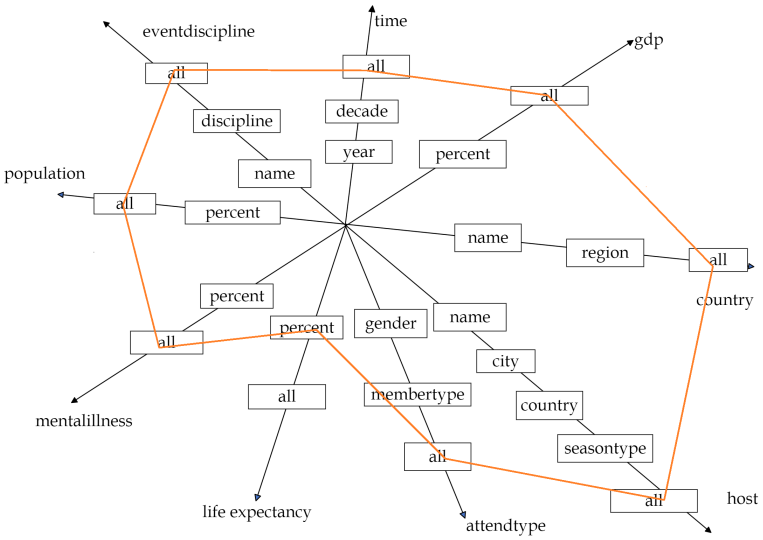


Figure 12: Query 11

1.2.4.12 Query 12 The query is: How do the continents compare in terms of their Olympic medal contributions by decades?

This can assist the IOC in understanding global sports development patterns and may influence where to focus their advertising or collaboration for increasing the influence.

The StarNet footprint should includes decade and region. (See Fig 13)

1.2.4.13 Query 13 The query is: What are the gold medal statistics for EU countries in the Olympic Games?

The EU might use this data to evaluate the success of member countries in international sports and possibly coordinate sports funding and training programs more effectively.

The StarNet footprint should includes country names to show countries in EU. (See Fig 14)

1.2.4.14 Query 14 The query is: How have EU gold medal wins varied by decade in the Olympic Games?

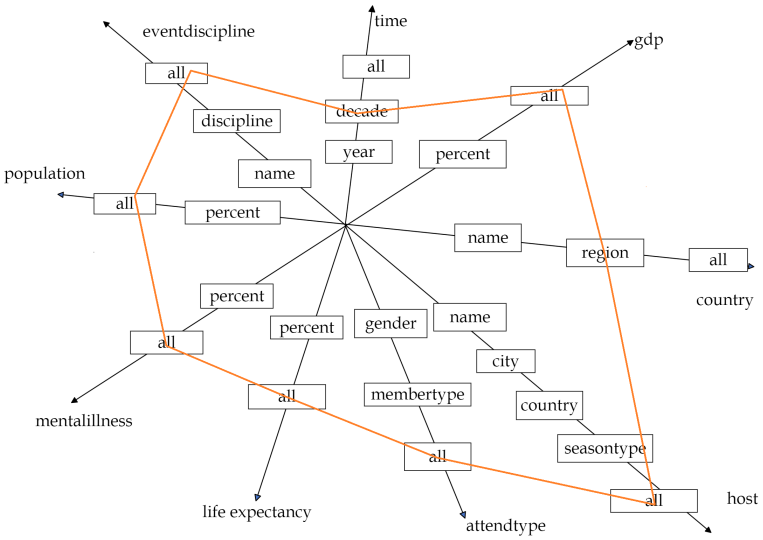


Figure 13: Query 12

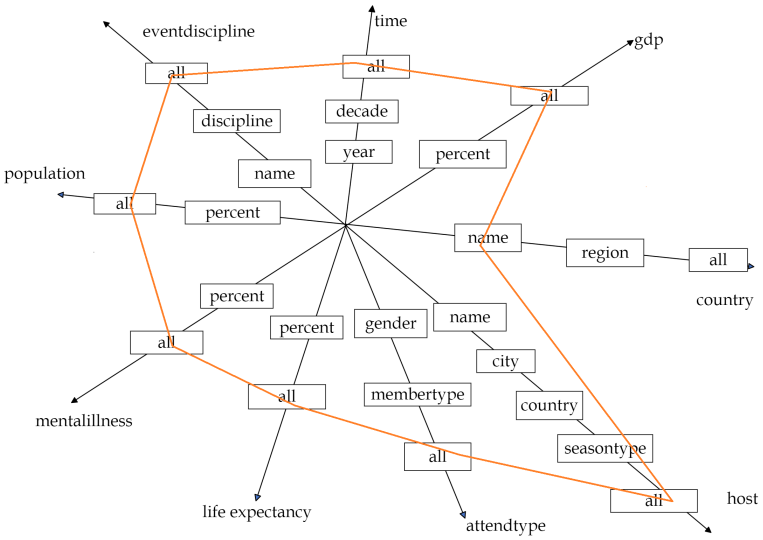


Figure 14: Query 13

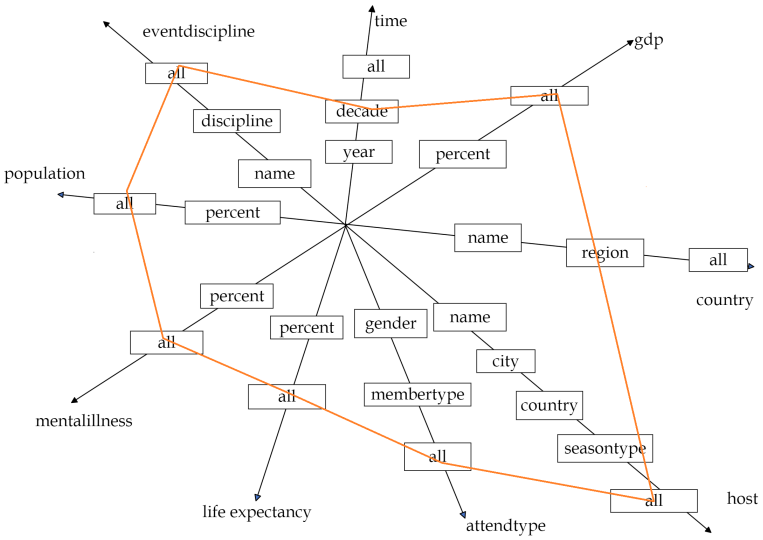


Figure 15: Query 14

This could provide insights into the effectiveness of EU sports initiatives and the impact of historical events on sports performance.

The StarNet footprint should includes decade and region. (See Fig 15)

1.2.4.15 Query 15 The query is: How does the gender distribution of Olympic medals compare between North America and Europe?

This question could help the EU understand gender equality in sports within these two countries.

The StarNet footprint should includes gender and region. (See Fig 16)

1.2.5 Database Schema

From this star schema design(Fig 17), OLAP become available for the analysts. For the relation of keys, only time dimension used natural key, all of the rest dimensions are using surrogate key because the data inside of time dimension does not change over time.

1.2.5.1 Dimensional Design Details Keys: From this star schema design, OLAP become available for the analysts. For the relation of keys, only time

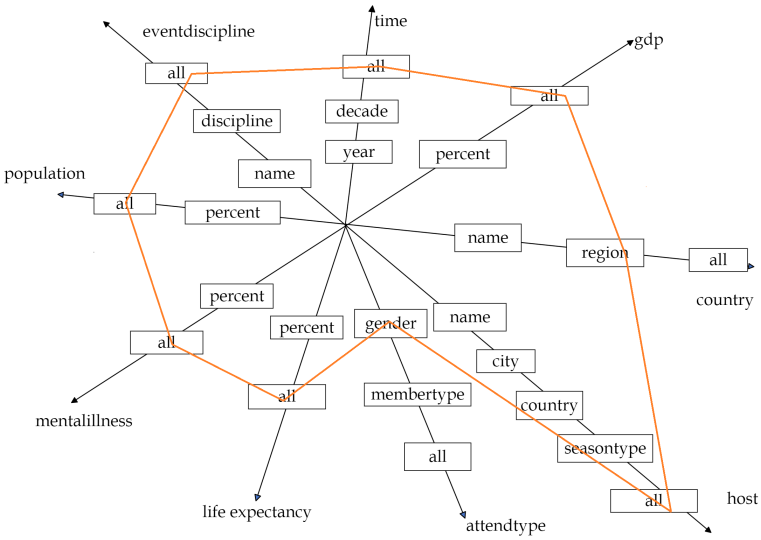


Figure 16: Query 15

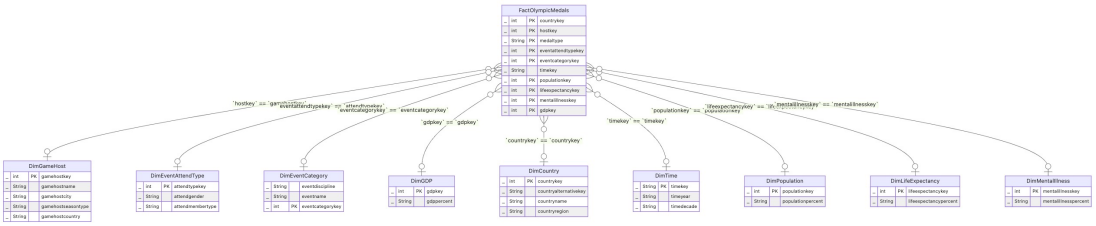


Figure 17: Star Schema

dimension used natural key, all of the rest dimensions are using surrogate key because the data inside of time dimension does not change over time. [1] Indices Dimensions: In order to put the data of GDP(to solve economic queries), life expectancy and mental illness(to solve health related queries) into dimensions, they are settled into different dimensional tables by setting group hierarchies.

Measures: There is only one measure inside of fact table - medaltype which contains the status of the medal. It would be used inside of PowerBI later for the measures of Olympic performance. Besides, filters would be explained and introduced in the following chapters. About the amount of dimensions: The data warehouse is still at it's best simplicity. The indices dimensions can not be put into a indice dimension because they can not be set into a hierarchy properly and it does not disobey the Kimball's dimensional modeling rule. It does not exceed 25 dimensions as centipede fact table.[1]

2 ETL

The structure of the ETL process of this data warehouse project would be csv file to OLTP server to OLAP, which means the data would be processed with ETL twice. Extracting data from OLTP database simulates the process of getting the data online. According to the provided data, it obviously has no performance problems because the data set is not that big. The data transformation all based on the csv file is also available because csv file is also a kind of data source and it may comes from OLTP servers. In this case, just simply change the tables from OLTP server to the parameters containing data frames from the initial cleaning would also be a nice choice. Here is only for showing how to extract the data from server. Ignoring this step could also be better for performance.(for this data warehouse the difference could almost be ignored) Also delete the foreign keys improves for the performance as when if the performance matters. [2] During the ETL process and OLAP design, Data Grip(PostgreSQL), jupyter lab(Python, Atoti) and PowerBI(hierarchies, data cube and measures) will be used as tools processing the data. Only the design, table figures would be shown and the explainations during the ETL process would be listed below. For further details, please refer to the

jupyter file for ETL process.

2.1 Initial cleaning and Data View

2.1.1 Path Setting

For the initial cleaning, the file paths would be put into a variable called base path by using pathlib package. The reason to put it into a variable is because the jupyter lab's working directory is set up on docker. For most of the users would be setting up working directory into project 1 folder, if they aim to use a different IDE, simply put the data address here will not be available for them. After setting the file path, a few lines of the code would examine whether the path exists, which might be useful to check if the working directory and data address is correct.

2.1.2 Initial cleaning

For the initial cleaning, it aims to clean the data to fit in the OLTP server and maintain the original shape by files. After having a glance of the data, the problems need to solve are:

1. Some titles are too long to put into the OLTP server. If the title name exceeds 64 characters, the data below would be missing.
2. Several column are not in used and it takes the storage.
3. Some of the table should be melted and some of them could be turned into pivot table for deciding the further transformation.
4. Some of the missing data could be initially turned into null for further processing.

The instructions of jupyter file would be explained table by table.

2.1.2.1 Economic Data.csv Although finally 2020_economy.csv is not in use, the information is not acquired at this stage. The data will still be cleaned for the initial analysis for the use to OLAP server. Firstly, the missing data with ".." was replaced by null, and the long title name changed into a short version.

And then drop the data containing year information because it only contains the data of 2020, which could be simply informed by the data name.

2.1.2.2 mental-illness.csv For the mental illness data, a pivot table was created from the original data file for viewing the data. The data would be cleaned seperately. Both of the tables would aggregates the different kinds of mental illnesses. The reason for this could be: No need for representing the current mental status in countries by different kinds of illnesses. An index would be created for pivot table as indentifier, and the entity column would be renamed. For the table for processing, only the title name would be changed.

2.1.2.3 Global Population.csv The population table would be melted, but the pivot table still in need. The missing data with "no data" will be simply changed into null, and additionally, the table would be melted for processing. And several of the country name would be changed for the merge of tables.

2.1.2.4 life-expectancy.csv The data before 1950 of life expectancy file lack of completeness. In order to view the data, a pivot table is needed. After deletion of data before 1950 and made into pivot table, melt into it's original form.

2.1.2.5 list-of-countries_areas-by-continent-2024.csv No need for cleaning at this stage, only changed the name.

2.1.2.6 olympic_hosts.csv From this file the host city and year could be generated by seperating game_slug column. Reformat the time data and rename the title into more reasonable ones.

2.1.2.7 olympic_medals.csv Repeat the seperation of game_slug is also available for this table. Besides, the information of athletes url and personal infomation won't be included because it consumes memory and contains no useful data for the queries. And it will not potentially useful for further use. After editting by the instructions above, change the name and several data which could be more reasonable to understand.

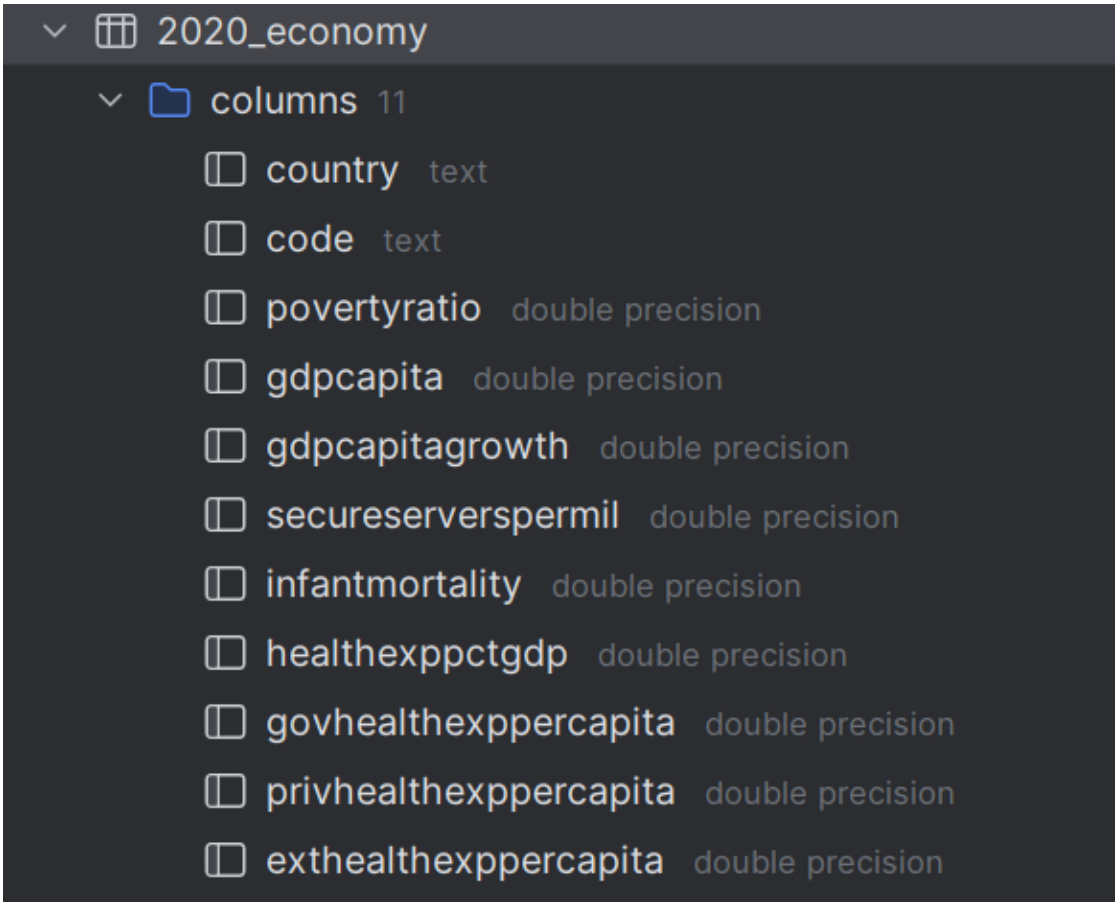


Figure 18: 2020 Economy

2.1.2.8 API_NY.GDP.MKTP.CD_DS2_en_csv_v2_26.csv This code snippet handles external data with long indices by rounding them as decimals. Delete the columns which are not necessary.

2.2 Read OLTP Server Data for Transformation

2.2.1 Read Data

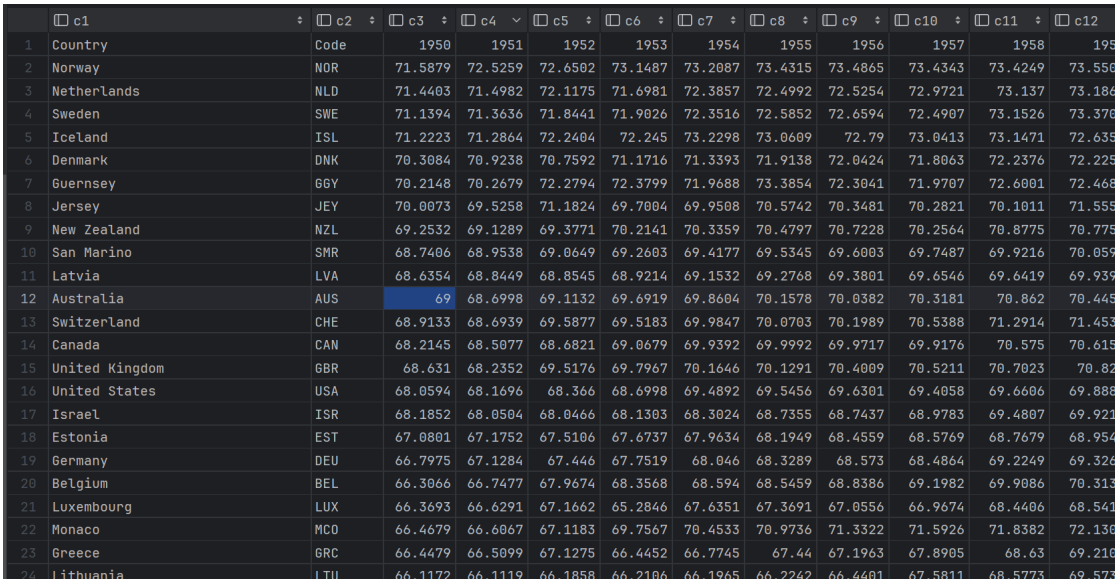
After the initial processing of the data, 11 tables from OLTP server are generated. The figures would be put in following pages. This part will only for explaining the data with more columns and also the pivot tables. 2020_economy table contains 11 columns which contains the economic data for 2020. Although it is abandoned,

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
1	Population (Millions of people)	1980	1981	1982	1983	1984	1985	1986	1987	1988
2	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
3	Afghanistan	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
4	Albania	2.672	2.726	2.784	2.844	2.904	2.965	3.023	3.084	3.142
5	Algeria	18.666	19.246	19.864	20.516	21.175	22.2	22.8	23.4	24.1
6	Andorra	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
7	Angola	8.272	8.495	8.72	8.948	9.185	10.35	10.646	10.918	11.214
8	Antigua and Barbuda	0.068	0.068	0.067	0.066	0.065	0.064	0.063	0.062	0.061
9	Argentina	27.95	28.45	28.93	29.34	29.84	30.35	30.74	31.09	31.47
10	Armenia	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
11	Aruba	<null>	<null>	<null>	<null>	<null>	<null>	0.061	0.06	0.061
12	Australia	14.802	15.039	15.289	15.483	15.677	15.901	16.139	16.395	16.687
13	Austria	7.54	7.556	7.565	7.543	7.544	7.549	7.557	7.567	7.576
14	Azerbaijan	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
15	Bahamas, The	0.211	0.215	0.22	0.225	0.23	0.234	0.238	0.242	0.247
16	Bahrain	0.353	0.367	0.38	0.393	0.406	0.42	0.435	0.45	0.467
17	Bangladesh	79.639	81.768	83.932	86.142	88.417	90.764	93.188	95.671	98.186
18	Barbados	0.252	0.253	0.254	0.255	0.255	0.256	0.257	0.258	0.259
19	Belarus	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
20	Belgium	9.855	9.863	9.855	9.858	9.853	9.858	9.859	9.865	9.876
21	Belize	0.15	0.15	0.153	0.157	0.162	0.166	0.171	0.176	0.179
22	Benin	3.75	3.88	3.999	4.129	4.259	4.4	4.541	4.682	4.833
23	Bhutan	0.408	0.419	0.43	0.44	0.451	0.464	0.479	0.495	0.511
24	Bolivia	5.161	5.587	5.703	5.822	5.944	6.068	6.195	6.324	6.456
25	Bosnia and Herzegovina	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
26	Botswana	0.963	0.981	1.017	1.055	1.093	1.134	1.176	1.219	1.263
27	Brazil	121.151	123.858	126.589	129.335	132.094	134.858	137.592	140.264	142.87
28	Brunei Darussalam	<null>	<null>	<null>	<null>	<null>	0.222	0.226	0.234	0.241
29	Bulgaria	8.808	8.834	8.859	8.881	8.897	8.903	8.901	8.889	8.865
30	Burkina Faso	7.045	7.212	7.391	7.58	7.776	7.979	8.208	8.435	8.664
31	Burundi	4.09	4.21	4.334	4.461	4.591	4.726	4.864	5.007	5.154
32	Cabo Verde	0.289	0.294	0.299	0.305	0.312	0.318	0.325	0.332	0.339
33	Cambodia	6.718	6.775	6.945	7.196	7.475	7.743	7.99	8.228	8.467

Figure 19: Population

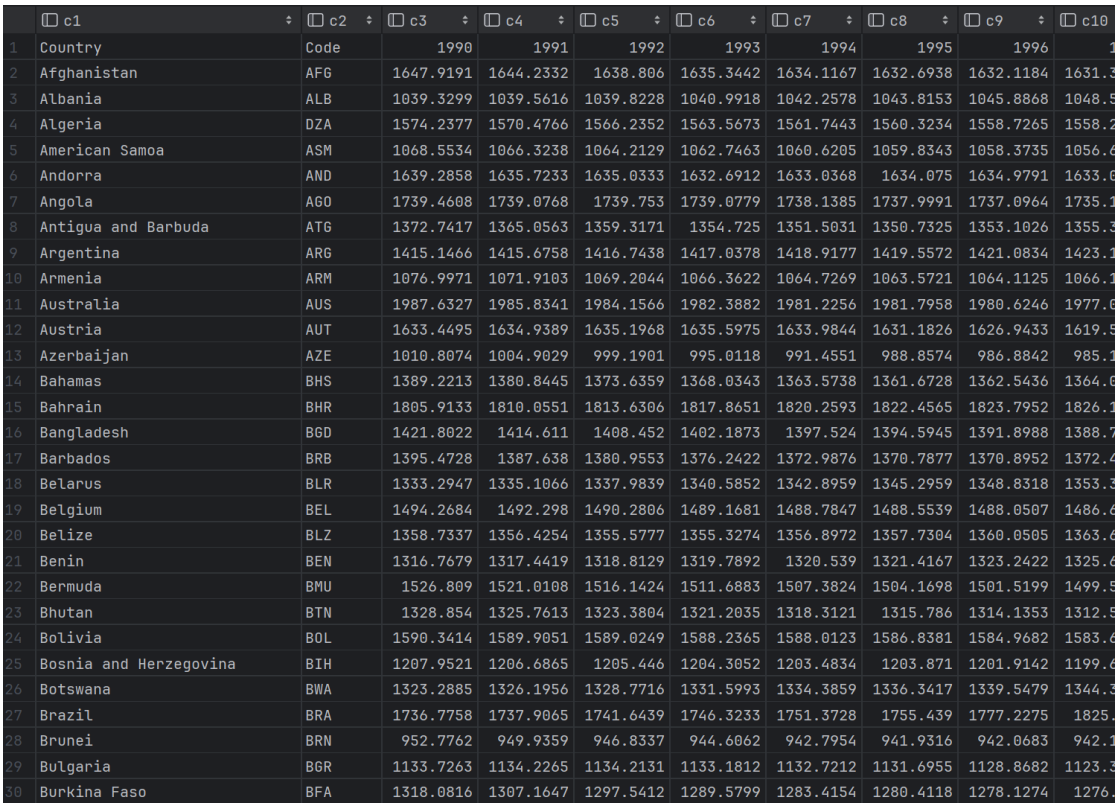
most of the information seems could also related to health but it is actually hard to relate to mental health data. The data is huge and the picture and show comprehensive itself so only the column names would be shown here. (see fig 18)

Population table (Fig 19), life expectancy table (Fig 20), mentall illness table (Fig 21) share the data in different range of time. In future processing, they would be put into the same time scale, and the null data left in fact table will not be edited and the data in dimensional table would be completed.



	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12
1	Country	Code	1950	1951	1952	1953	1954	1955	1956	1957	1958	195
2	Norway	NOR	71.5879	72.5259	72.6502	73.1487	73.2087	73.4315	73.4865	73.4343	73.4249	73.550
3	Netherlands	NLD	71.4403	71.4982	72.1175	71.6981	72.3857	72.4992	72.5254	72.9721	73.137	73.186
4	Sweden	SWE	71.1394	71.3636	71.8441	71.9026	72.3516	72.5852	72.6594	72.4907	73.1526	73.370
5	Iceland	ISL	71.2223	71.2864	72.2404	72.245	73.2298	73.0609	72.79	73.0413	73.1471	72.639
6	Denmark	DNK	70.3084	70.9238	70.7592	71.1716	71.3393	71.9138	72.0424	71.8063	72.2376	72.225
7	Guernsey	GGY	70.2148	70.2679	72.2794	72.3799	71.9688	73.3854	72.3041	71.9707	72.6001	72.468
8	Jersey	JEY	70.0073	69.5258	71.1824	69.7004	69.9508	70.5742	70.3481	70.2821	70.1011	71.559
9	New Zealand	NZL	69.2532	69.1289	69.3771	70.2141	70.3359	70.4797	70.7228	70.2564	70.8775	70.775
10	San Marino	SMR	68.7406	68.9538	69.0649	69.2603	69.4177	69.5345	69.6003	69.7487	69.9216	70.059
11	Latvia	LVA	68.6354	68.8449	68.8545	68.9214	69.1532	69.2768	69.3801	69.6546	69.6419	69.939
12	Australia	AUS	69	68.6998	69.1132	69.6919	69.8604	70.1578	70.0382	70.3181	70.862	70.445
13	Switzerland	CHE	68.9133	68.6939	69.5877	69.5183	69.9847	70.0703	70.1989	70.5388	71.2914	71.453
14	Canada	CAN	68.2145	68.5077	68.6821	69.0679	69.9392	69.9992	69.9717	69.9176	70.575	70.615
15	United Kingdom	GBR	68.631	68.2352	69.5176	69.7967	70.1646	70.1291	70.4009	70.5211	70.7023	70.82
16	United States	USA	68.0594	68.1696	68.366	68.6998	69.4892	69.5456	69.6301	69.4058	69.6606	69.888
17	Israel	ISR	68.1852	68.0504	68.0466	68.1303	68.3024	68.7355	68.7437	68.9783	69.4807	69.921
18	Estonia	EST	67.0801	67.1752	67.5106	67.6737	67.9634	68.1949	68.4559	68.5769	68.7679	68.954
19	Germany	DEU	66.7975	67.1284	67.446	67.7519	68.046	68.3289	68.573	68.4864	69.2249	69.326
20	Belgium	BEL	66.3066	66.7477	67.9674	68.3568	68.594	68.5459	68.8386	69.1982	69.9086	70.313
21	Luxembourg	LUX	66.3693	66.6291	67.1662	65.2846	67.6351	67.3691	67.0556	66.9674	68.4406	68.541
22	Monaco	MCO	66.4679	66.6067	67.1183	69.7567	70.4533	70.9736	71.3322	71.5926	71.8382	72.130
23	Greece	GRC	66.4479	66.5099	67.1275	66.4452	66.7745	67.44	67.1963	67.8905	68.63	69.210
24	Lithuania	LTU	66.1172	66.1119	66.1858	66.2106	66.1965	66.2242	66.4401	67.5811	68.5773	69.573

Figure 20: Life Expectancy



	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
1	Country	Code	1990	1991	1992	1993	1994	1995	1996	1
2	Afghanistan	AFG	1647.9191	1644.2332	1638.806	1635.3442	1634.1167	1632.6938	1632.1184	1631.3
3	Albania	ALB	1039.3299	1039.5616	1039.8228	1040.9918	1042.2578	1043.8153	1045.8868	1048.5
4	Algeria	DZA	1574.2377	1570.4766	1566.2352	1563.5673	1561.7443	1560.3234	1558.7265	1558.2
5	American Samoa	ASM	1068.5534	1066.3238	1064.2129	1062.7463	1060.6205	1059.8343	1058.3735	1056.6
6	Andorra	AND	1639.2858	1635.7233	1635.0333	1632.6912	1633.0368	1634.075	1634.9791	1633.0
7	Angola	AGO	1739.4608	1739.0768	1739.753	1739.0779	1738.1385	1737.9991	1737.0964	1735.1
8	Antigua and Barbuda	ATG	1372.7417	1365.0563	1359.3171	1354.725	1351.5031	1350.7325	1353.1026	1355.3
9	Argentina	ARG	1415.1466	1415.6758	1416.7438	1417.0378	1418.9177	1419.5572	1421.0834	1423.1
10	Armenia	ARM	1076.9971	1071.9103	1069.2044	1066.3622	1064.7269	1063.5721	1064.1125	1066.1
11	Australia	AUS	1987.6327	1985.8341	1984.1566	1982.3882	1981.2256	1981.7958	1980.6246	1977.0
12	Austria	AUT	1633.4495	1634.9389	1635.1968	1635.5975	1633.9844	1631.1826	1626.9433	1619.9
13	Azerbaijan	AZE	1010.8074	1004.9029	999.1901	995.0118	991.4551	988.8574	986.8842	985.1
14	Bahamas	BHS	1389.2213	1380.8445	1373.6359	1368.0343	1363.5738	1361.6728	1362.5436	1364.0
15	Bahrain	BHR	1805.9133	1810.0551	1813.6306	1817.8651	1820.2593	1822.4565	1823.7952	1826.1
16	Bangladesh	BGD	1421.8022	1414.611	1408.452	1402.1873	1397.524	1394.5945	1391.8988	1388.7
17	Barbados	BRB	1395.4728	1387.638	1380.9553	1376.2422	1372.9876	1370.7877	1370.8952	1372.4
18	Belarus	BLR	1333.2947	1335.1066	1337.9839	1340.5852	1342.8959	1345.2959	1348.8318	1353.3
19	Belgium	BEL	1494.2684	1492.298	1490.2806	1489.1681	1488.7847	1488.5539	1488.0507	1486.6
20	Belize	BLZ	1358.7337	1356.4254	1355.5777	1355.3274	1356.8972	1357.7304	1360.0505	1363.6
21	Benin	BEN	1316.7679	1317.4419	1318.8129	1319.7892	1320.539	1321.4167	1323.2422	1325.6
22	Bermuda	BMU	1526.809	1521.0108	1516.1424	1511.6883	1507.3824	1504.1698	1501.5199	1499.5
23	Bhutan	BTN	1328.854	1325.7613	1323.3804	1321.2035	1318.3121	1315.786	1314.1353	1312.5
24	Bolivia	BOL	1590.3414	1589.9051	1589.0249	1588.2365	1588.0123	1586.8381	1584.9682	1583.6
25	Bosnia and Herzegovina	BIH	1207.9521	1206.6865	1205.446	1204.3052	1203.4834	1203.871	1201.9142	1199.6
26	Botswana	BWA	1323.2885	1326.1956	1328.7716	1331.5993	1334.3859	1336.3417	1339.5479	1344.3
27	Brazil	BRA	1736.7758	1737.9065	1741.6439	1746.3233	1751.3728	1755.439	1777.2275	1825.
28	Brunei	BRN	952.7762	949.9359	946.8337	944.6062	942.7954	941.9316	942.0683	942.1
29	Bulgaria	BGR	1133.7263	1134.2265	1134.2131	1133.1812	1132.7212	1131.6955	1128.8682	1123.3
30	Burkina Faso	BFA	1318.0816	1307.1647	1297.5412	1289.5799	1283.4154	1280.4118	1278.1274	1276.

Figure 21: Mental Illness

2.2.2 Data Transformation

After understanding the information of the tables, further processing of the data would be utilized. The process is a little bit complicated and only the brief would be listed below.

1. 2020economy: Abandoned
2. location: Merge the country code as alternative key from mental illness table. Update the keys of several countries and create surrogate keys for the table. Add the countries which are not exists now and also missing from the table manually, which is known by the missing value check later in the medal table building.
3. host: Create surrogate keys and rename the country name contains different names. The reason of the historical country changed to the country exists now because the cities are actually now in those countries.
4. category: Divided from medal table by set unique and create surrogate keys for that.
5. attend: Also from medal table, but edited the data of gender to fulfil the hierarchy of this table.
6. population, gdp, mentalillness, lifeexpectancy: Created manually to percentages divided by 10 with surrogate keys.
7. time: From the 1869 to 2023 which contains all the years of historical Olympic games, and add decades column for larger scale queries, then creates its natural key.
8. medal: As fact table, the country code was contructed as ISO country code which is not sufficient for Olympics and will convert it into IOC country codes. Merged the tables as dimensions for its keys. The data of indices would be set grouped into percentages.

The details are described by the comments on ipynb file. After processing like above. 10 tables are generated. The tables would be shown below.(See Fig 22-31)

	attendtypekey	attendgender	attendmembertype
1	1	TeamMixed	Team
2	2	TeamWomen	Team
3	3	TeamMen	Team
4	4	SingleMen	Single
5	5	SingleWomen	Single
6	6	TeamOpen	Team
7	7	SingleOpen	Single

Figure 22: dimeventattendtype

	countrykey	countryalternativekey	countryname	countryregion
1	1	IND	India	Asia
2	2	CHN	China	Asia
3	3	USA	United States	North America
4	4	IDN	Indonesia	Asia
5	5	PAK	Pakistan	Asia
6	6	NGA	Nigeria	Africa
7	7	BRA	Brazil	South America
8	8	BGD	Bangladesh	Asia
9	9	RUS	Russia	Europe
10	10	ETH	Ethiopia	Africa
11	11	MEX	Mexico	North America
12	12	JPN	Japan	Asia
13	13	PHL	Philippines	Asia
14	14	EGY	Egypt	Africa
15	15	COD	DR Congo	Africa
16	16	VNM	Vietnam	Asia
17	17	IRN	Iran	Asia
18	18	TUR	Turkey	Asia
19	19	DEU	Germany	Europe

Figure 23: dimcountry

2.3 PostSQL Build

The data base would created based on the design of dimensional design. All of the primary keys would be surrogate keys except dimtime. And every column of fact table would have a foreign key constraint to the related dimensional table. The data would be stored into the olympicsOLAP database. The database creation only needs to import the csv file to OLTP database, and manually create the tables inside of OLAP database. For DataGrip users, there is no need to write SQL queries inside of python manually for creating tables. The SQL scripts would be provided in order inside of the zip file.

	eventdiscipline	eventname	eventcategorykey
1	Curling	Mixed Doubles	1
2	Curling	Women	2
3	Curling	Men	3
4	Freestyle Skiing	Men's Moguls	4
5	Freestyle Skiing	Men's Freeski Halfpipe	5
6	Freestyle Skiing	Men's Freeski Big Air	6
7	Freestyle Skiing	Men's Ski Cross	7
8	Freestyle Skiing	Women's Freeski Big Air	8
9	Freestyle Skiing	Women's Moguls	9
10	Freestyle Skiing	Mixed Team Aerials	10
11	Freestyle Skiing	Women's Ski Cross	11
12	Freestyle Skiing	Men's Aerials	12
13	Freestyle Skiing	Women's Aerials	13
14	Freestyle Skiing	Women's Freeski Halfpipe	14
15	Freestyle Skiing	Women's Freeski Slopestyle	15

Figure 24: dimeventcategory

	gdpkey	gdppercent
1	1	0
2	2	10
3	3	20
4	4	30
5	5	40
6	6	50
7	7	60
8	8	70
9	9	80
10	10	90
11	11	100

Figure 25: dimgdp

	gamehostkey	gamehostname	gamehostcity	gamehostseasontype	gamehostcountry
1	1	Beijing 2022	Beijing	Winter	China
2	2	Tokyo 2020	Tokyo	Summer	Japan
3	3	PyeongChang 2018	Pyeongchang	Winter	North Korea
4	4	Rio 2016	Rio	Summer	Brazil
5	5	Sochi 2014	Sochi	Winter	Russia
6	6	London 2012	London	Summer	United Kingdom
7	7	Vancouver 2010	Vancouver	Winter	Canada
8	8	Beijing 2008	Beijing	Summer	China
9	9	Turin 2006	Turin	Winter	Italy
10	10	Athens 2004	Athens	Summer	Greece
11	11	Salt Lake City 2002	Salt Lake City	Winter	United States
12	12	Sydney 2000	Sydney	Summer	Australia
13	13	Nagano 1998	Nagano	Winter	Japan
14	14	Atlanta 1996	Atlanta	Summer	United States
15	15	Lillehammer 1994	Lillehammer	Winter	Norway
16	16	Barcelona 1992	Barcelona	Summer	Spain
17	17	Albertville 1992	Albertville	Winter	France

Figure 26: dimgamehost

	lifeexpectancykey	lifeexpectancypercent
1	1	0
2	2	10
3	3	20
4	4	30
5	5	40
6	6	50
7	7	60
8	8	70
9	9	80
10	10	90
11	11	100

Figure 27: dimlifeexpectancy

	mentalillnesskey	mentalillnesspercent
1	1	0
2	2	10
3	3	20
4	4	30
5	5	40
6	6	50
7	7	60
8	8	70
9	9	80
10	10	90
11	11	100

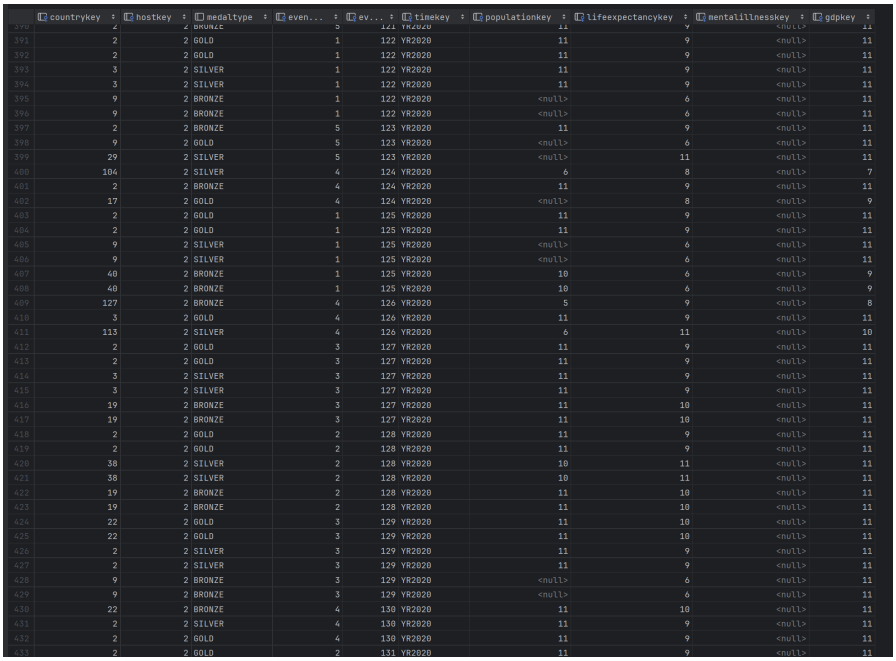
Figure 28: dimmentalillness

	populationkey	populationpercent
1	1	0
2	2	10
3	3	20
4	4	30
5	5	40
6	6	50
7	7	60
8	8	70
9	9	80
10	10	90
11	11	100

Figure 29: dimpopulation

	timekey	timeyear	timedecade
1	YR1896	1896	1890
2	YR1897	1897	1890
3	YR1898	1898	1890
4	YR1899	1899	1890
5	YR1900	1900	1900
6	YR1901	1901	1900
7	YR1902	1902	1900
8	YR1903	1903	1900
9	YR1904	1904	1900
10	YR1905	1905	1900
11	YR1906	1906	1900
12	YR1907	1907	1900
13	YR1908	1908	1900
14	YR1909	1909	1900
15	YR1910	1910	1910
16	YR1911	1911	1910

Figure 30: dimtime



	countrykey	hostkey	medaltype	even...	ev...	timekey	populationkey	lifeexpectancykey	mentalillnesskey	gdpkey
391	2	2	GOLD	1	122	YR2020	11	9	<null>	11
392	2	2	GOLD	1	122	YR2020	11	9	<null>	11
393	3	2	SILVER	1	122	YR2020	11	9	<null>	11
394	3	2	SILVER	1	122	YR2020	11	9	<null>	11
395	9	2	BRONZE	1	122	YR2020	<null>	6	<null>	11
396	9	2	BRONZE	1	122	YR2020	<null>	6	<null>	11
397	2	2	BRONZE	5	123	YR2020	11	9	<null>	11
398	9	2	GOLD	5	123	YR2020	<null>	6	<null>	11
399	29	2	SILVER	5	123	YR2020	<null>	11	<null>	11
400	104	2	SILVER	4	124	YR2020	6	8	<null>	7
401	2	2	BRONZE	4	124	YR2020	11	9	<null>	11
402	17	2	GOLD	4	124	YR2020	<null>	8	<null>	9
403	2	2	GOLD	1	125	YR2020	11	9	<null>	11
404	2	2	GOLD	1	125	YR2020	11	9	<null>	11
405	9	2	SILVER	1	125	YR2020	<null>	6	<null>	11
406	9	2	SILVER	1	125	YR2020	<null>	6	<null>	11
407	40	2	BRONZE	1	125	YR2020	10	6	<null>	9
408	40	2	BRONZE	1	125	YR2020	10	6	<null>	9
409	127	2	BRONZE	4	126	YR2020	5	9	<null>	8
410	3	2	GOLD	4	126	YR2020	11	9	<null>	11
411	113	2	SILVER	4	126	YR2020	6	11	<null>	10
412	2	2	GOLD	3	127	YR2020	11	9	<null>	11
413	2	2	GOLD	3	127	YR2020	11	9	<null>	11
414	3	2	SILVER	3	127	YR2020	11	9	<null>	11
415	3	2	SILVER	3	127	YR2020	11	9	<null>	11
416	19	2	BRONZE	3	127	YR2020	11	10	<null>	11
417	19	2	BRONZE	3	127	YR2020	11	10	<null>	11
418	2	2	GOLD	2	128	YR2020	11	9	<null>	11
419	2	2	GOLD	2	128	YR2020	11	9	<null>	11
420	38	2	SILVER	2	128	YR2020	10	11	<null>	11
421	38	2	SILVER	2	128	YR2020	10	11	<null>	11
422	19	2	BRONZE	2	128	YR2020	11	10	<null>	11
423	19	2	BRONZE	2	128	YR2020	11	10	<null>	11
424	22	2	GOLD	3	129	YR2020	11	10	<null>	11
425	22	2	GOLD	3	129	YR2020	11	10	<null>	11
426	2	2	SILVER	3	129	YR2020	11	9	<null>	11
427	2	2	SILVER	3	129	YR2020	11	9	<null>	11
428	9	2	BRONZE	3	129	YR2020	<null>	6	<null>	11
429	9	2	BRONZE	3	129	YR2020	<null>	6	<null>	11
430	22	2	BRONZE	4	130	YR2020	11	10	<null>	11
431	2	2	SILVER	4	130	YR2020	11	9	<null>	11
432	2	2	GOLD	4	130	YR2020	11	9	<null>	11
433	2	2	GOLD	2	131	YR2020	11	9	<null>	11

Figure 31: factolympicmedals

2.4 OLAP Design

2.4.1 data cube and Atoti measures and hierarchies

The data cube would be created by Atoti(PJ1 Atoti file) reading tables from PostgreSQL and then join them to create a data cube which could pre-compute the data in dimensions. Then add hierarchies (Fig 32)and measures(Fig 33).

In order to measure the achievements of olympic medals, two measures had been built by counting the quantities of gold medals and setting points for medal types.(Gold = 3, Silver = 2, Bronze = 1) The scores could represents the grades of the Olympic medals in an aggregated way.

The cube queries are used to first have a glance of the result of the data, which could also answer the queries. In order to minimize the understanding of data, the visualization follows.

After loading the data from PostgreSQL to PowerBI, The tables would be automatically linked together. Only several hierarchies and measures would be created

Figure 32: Hierarchies Atoti

```
{
  "Gold Medals Count": {
    "formatter": "INT[#,###]"
  },
  "Medal Score": {
    "formatter": "INT[#,###]"
  }
}
```

Figure 33: Measures Atoti

based on the clients' requirements.

2.4.2 Hierarchies inside of PowerBI

For the hierarchies for the graph, several of them could be simply created and adjusted by dragging them together. The hierarchies are created as below for the use of visualization.(See Fig 34-35)

2.4.3 Measures of PowerBI by DAX

The measures are defined by DAX with the same definition as Atoti version.(See Fig 36-37)

The design of these hierarchies and measures enables users to conduct roll-up and drill-down data analysis, greatly leveraging the advantages of dimension tables.

3 Visualization

3.1 Queries Answering

3.1.1 Query 1

In which ten disciplines do Australian athletes consistently achieve the highest performance?

By visualization, swmming, athletics, rowing, sailing, cycling track, canoe sprint, diving, equestrian eventing, hockey and shooting are the top 10 disciplines. (Fig 38) From the information above, Australian athletes are really good at aquatic sports. Besides, shooting, hockey, equestrian eventing potentially could become Australians' strength.

3.1.2 Query 2

Can we observe a trend in Australia's overall historical performance in the Olympic Games or winter games by years?

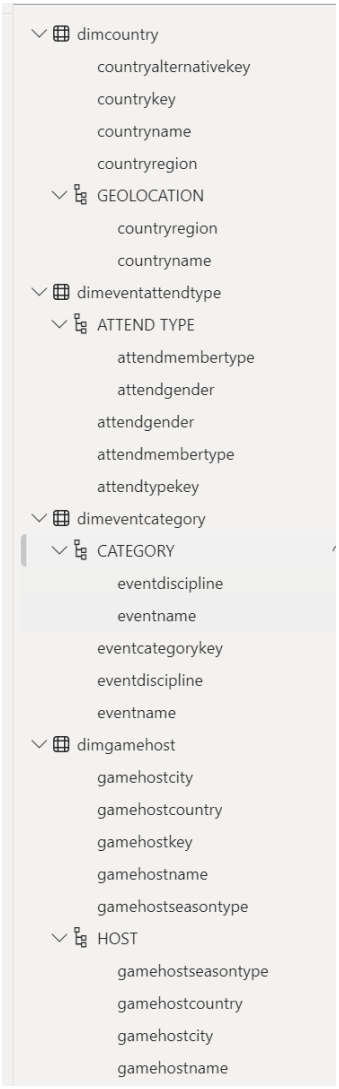


Figure 34: Hierarchies 1

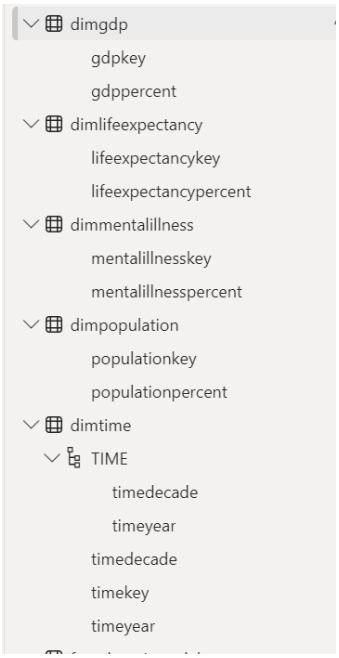


Figure 35: Hierarchies 2

```
Gold Medals Count = COUNTX(
    FILTER(
        factOlympicMedals,
        factOlympicMedals[MedalType] = "Gold"
    ),
    factOlympicMedals[MedalType]
)

```

Figure 36: Measures 1

```
Medal Score = SUMX(
    factOlympicMedals,
    SWITCH(
        factOlympicMedals[MedalType],
        "Gold", 3,
        "Silver", 2,
        "Bronze", 1,
        0
    )
)

```

Figure 37: Measures 2

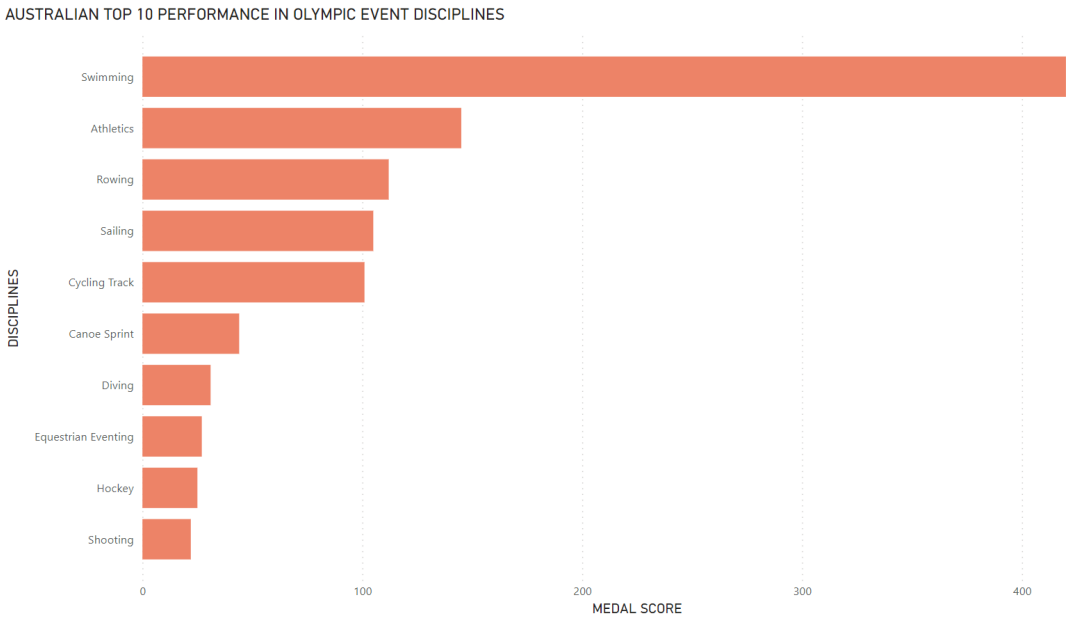


Figure 38: Query 1

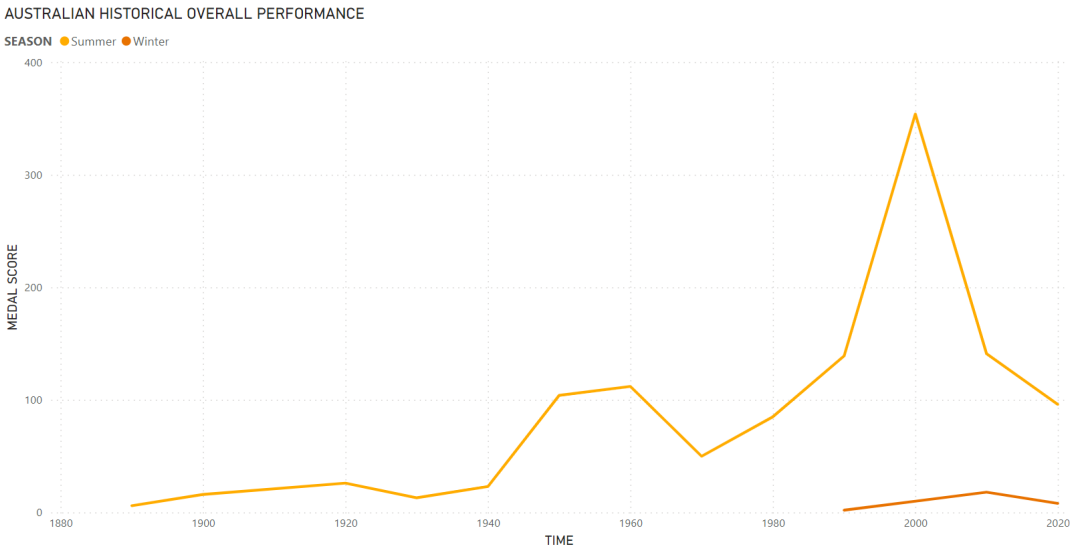


Figure 39: Query 2 Decade

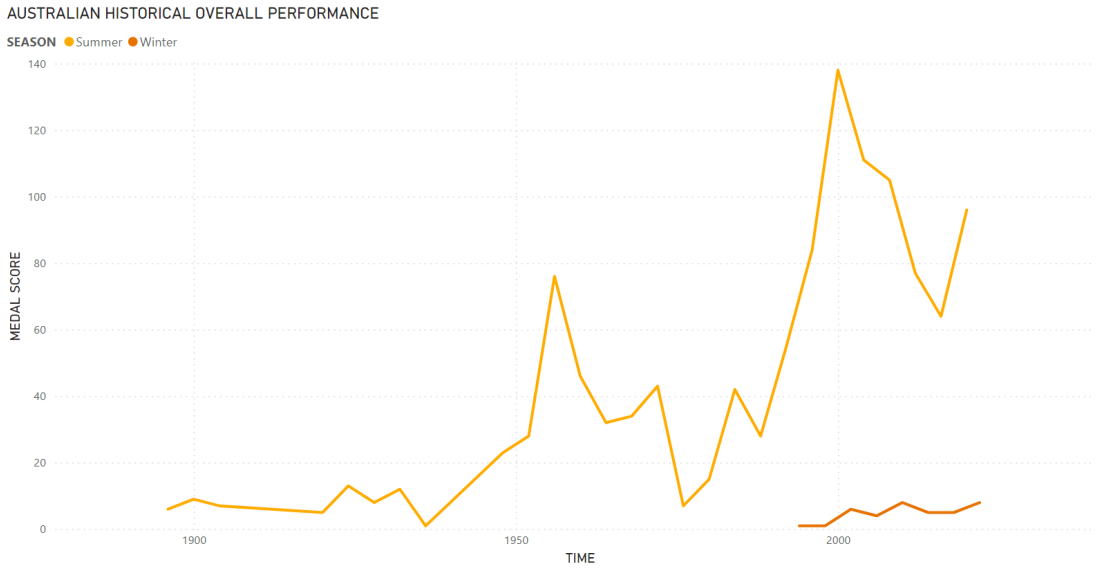


Figure 40: Query 2 Yearly

The line graphs presented illustrate the historical overall performance of Australia by decades (Figure 39) for roll-up analysis and years (Figure 40). Both graphs indicate a peak in performance during 2000, showing a dramatic increase and subsequent fall. The trend shows that the performance of Australia is becoming better but AOC might have to learn from the success from 2000.

3.1.3 Query 3

Could you provide a visualization that represents the distribution of genders across all Olympic events by years (Not including Open and Mixed)?

Figure 41 and 42 shows the growing trend of women events. Until recent years the Olympics almost achieved gender equality. There are still 4 percent difference between genders and it could be useful for IOC to identify how to balance for it.

3.1.4 Query 4

Is there a correlation between the mental well-being of athletes and the acquisition of gold medals?

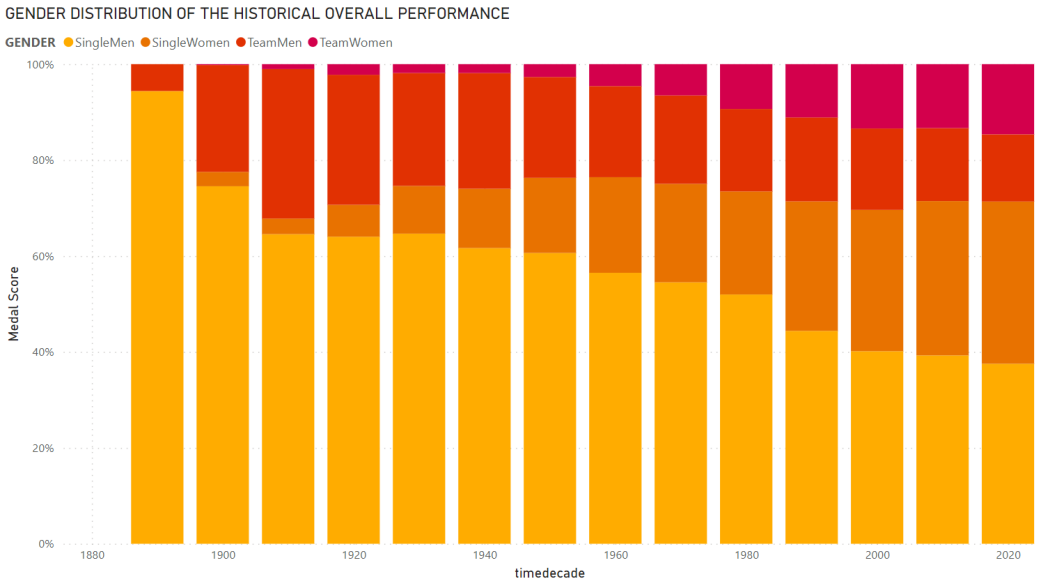


Figure 41: Query 3 Decade

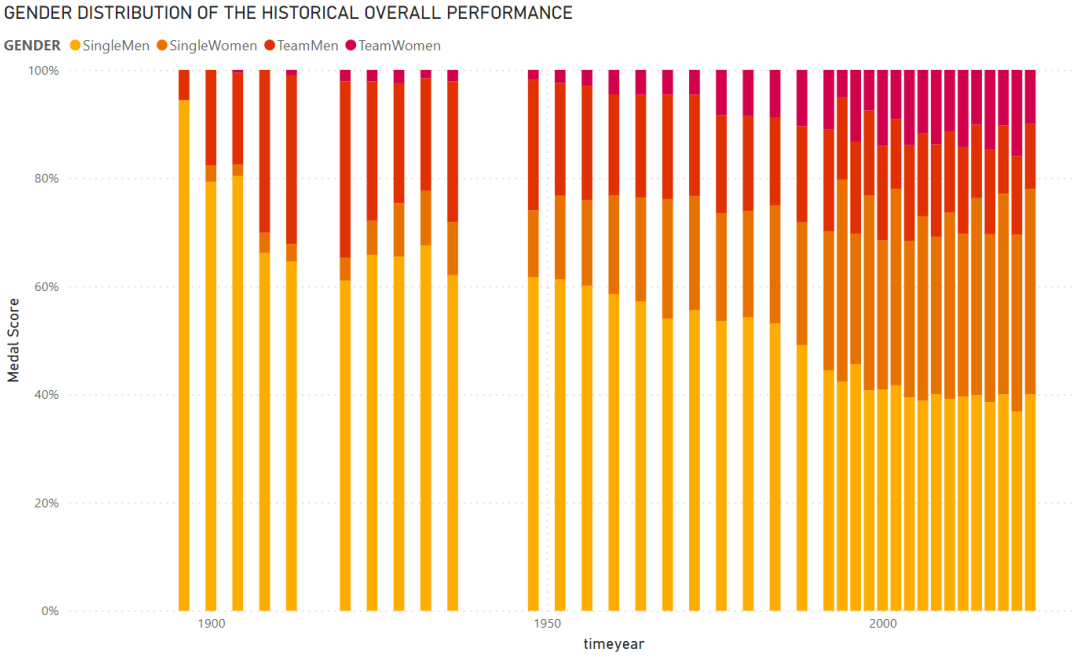


Figure 42: Query 3 Yearly

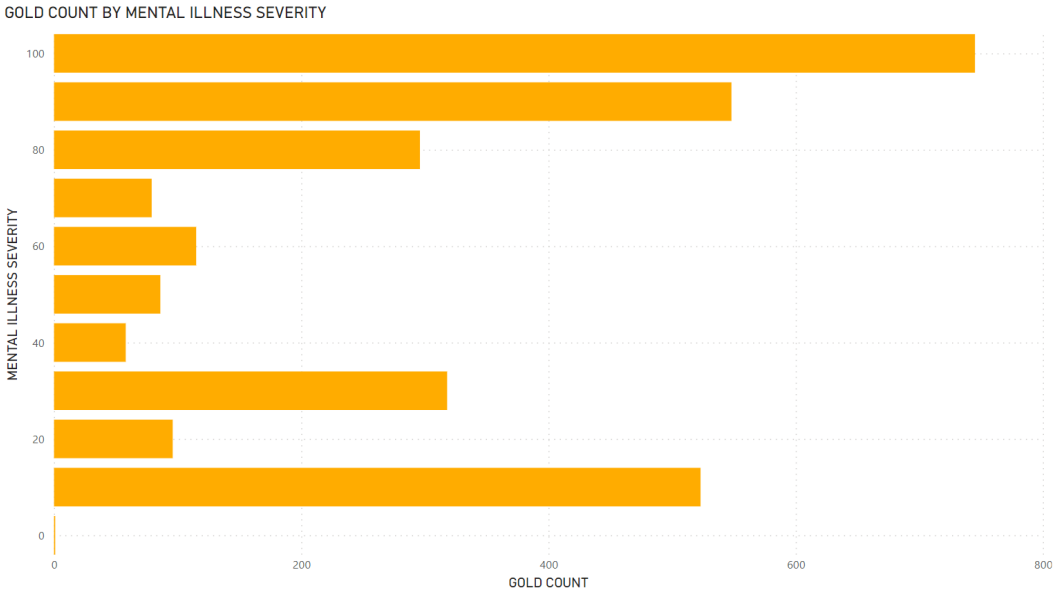


Figure 43: Query 4

From figure 43, the mental illness data is obviously related to gold count, which means they are in bimodal distribution. It means athletes from the worst or best mental illness severity country could be easier to earn gold count. This could be the information for Olympics to improve their fairness by eliminating this factor.

3.1.5 Query 5

What are the ten disciplines where European athletes tend to have the least success?

The data (Fig 44) shows skateboarding, rugby sevens, lacrosse, Jeu de Paume, cricket, trampoline gymnastics, cycling BMX freestyle, military patrol, basque pelota and sport climbing are the diciplines which Europeans not good at. EU may adjust their investment strategy focus on these 10 discipline to empower their Olympic team.

3.1.6 Query 6

How economy impress the performance of Olympics?

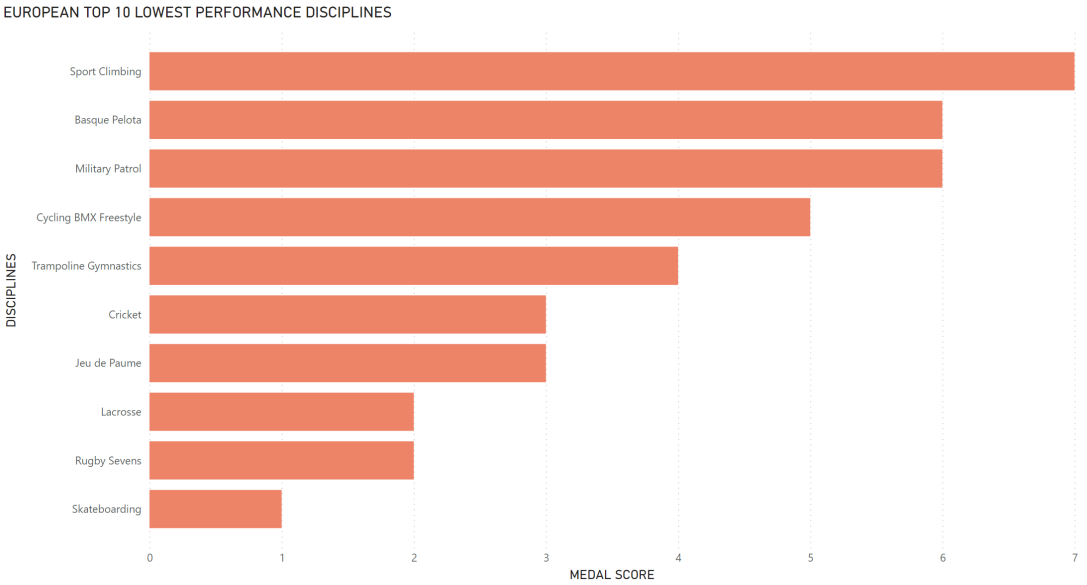


Figure 44: Query 5

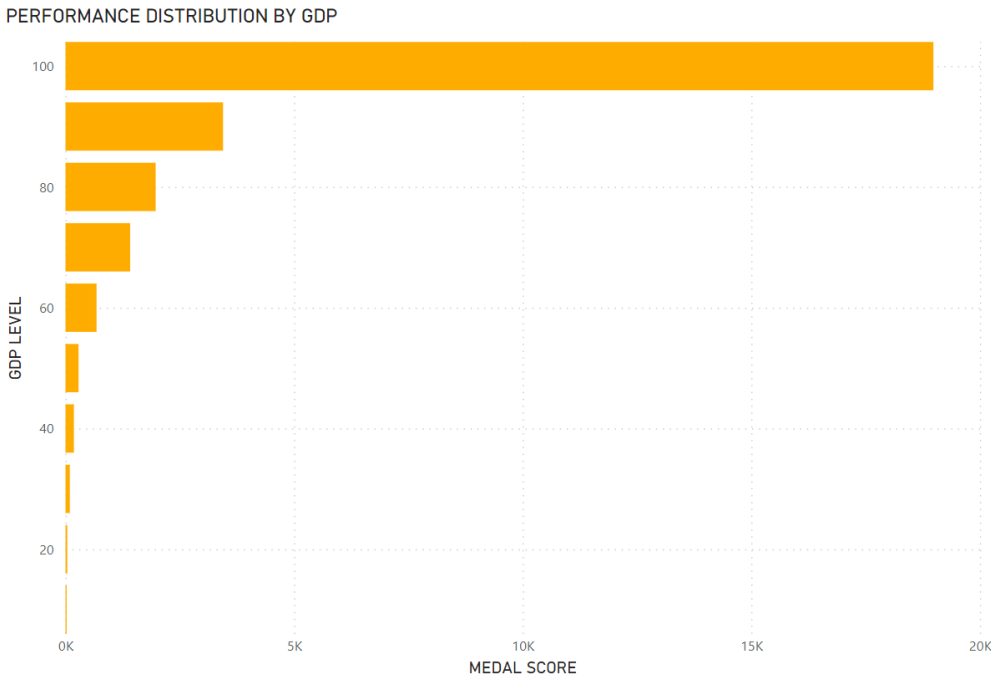


Figure 45: Query 6

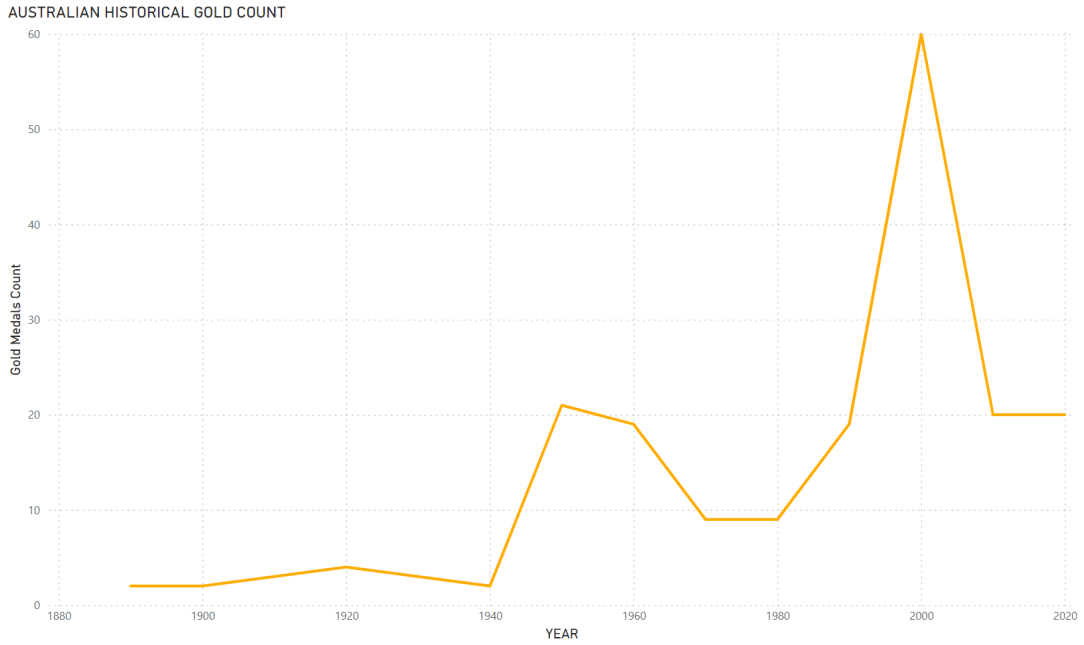


Figure 46: Query 7

The bar chart shows that gdp level is highly correlated to the Olympic performance. Which would be useful for EU to decide their investments by concerning the performance of European athletes.(Fig 45)

3.1.7 Query 7

How many gold medals has Australia won in each decade of the Olympic Games?

The charts hows that illustrate the historical gold medal count of Australia by decades. It shows almost the same trend as medal count and Australia may still need to investigate how they achieved that high in 2000s. (Fig 46)

3.1.8 Query 8

How do Australia's Olympic medal counts compare to those of New Zealand across all Olympic Games?

This graph shows that gender equality in Olympics is better than NZ in AU. Which means that Australia owns the leading place in the gender equality. (Fig 47)

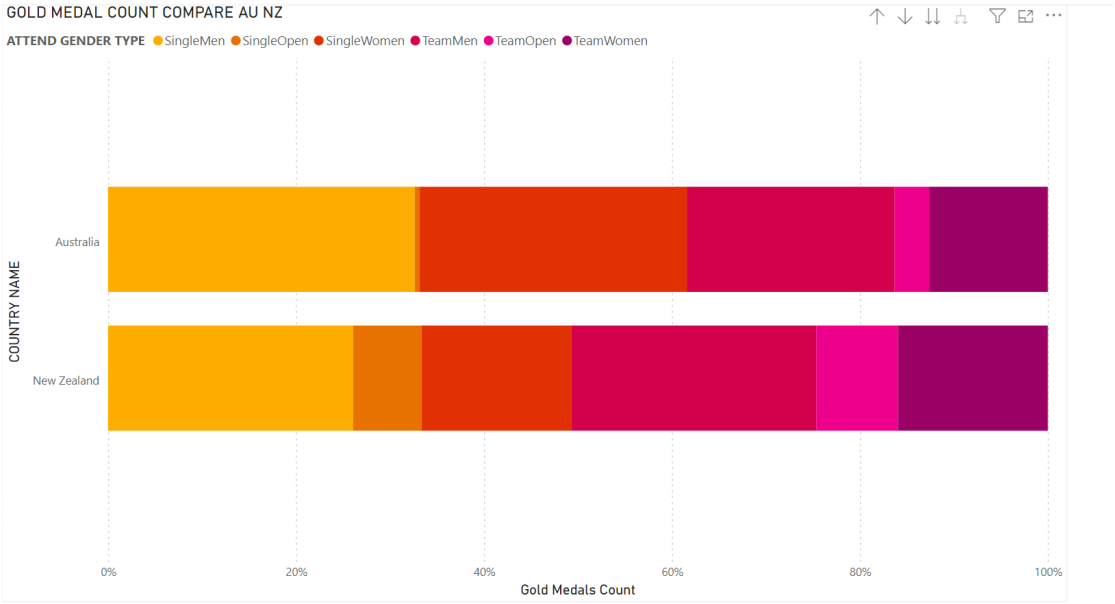


Figure 47: Query 8

3.1.9 Query 9

What is the contribution of Oceania countries in terms of Olympic medal?

This pie chart shows that Australia and NZ got most of the medals in Oceania. Australia still is the leader of Olympics in Oceania and this may be used to compare the data in later years. (Fig 48)

3.1.10 Query 10

Is there a correlation between a country's population size and its number of Olympic gold medals?

This shows that population is highly correlated to gold medal. However, there is a small hill peak at 60 percent which means the data may be affected by many other aspects. (Fig 49)

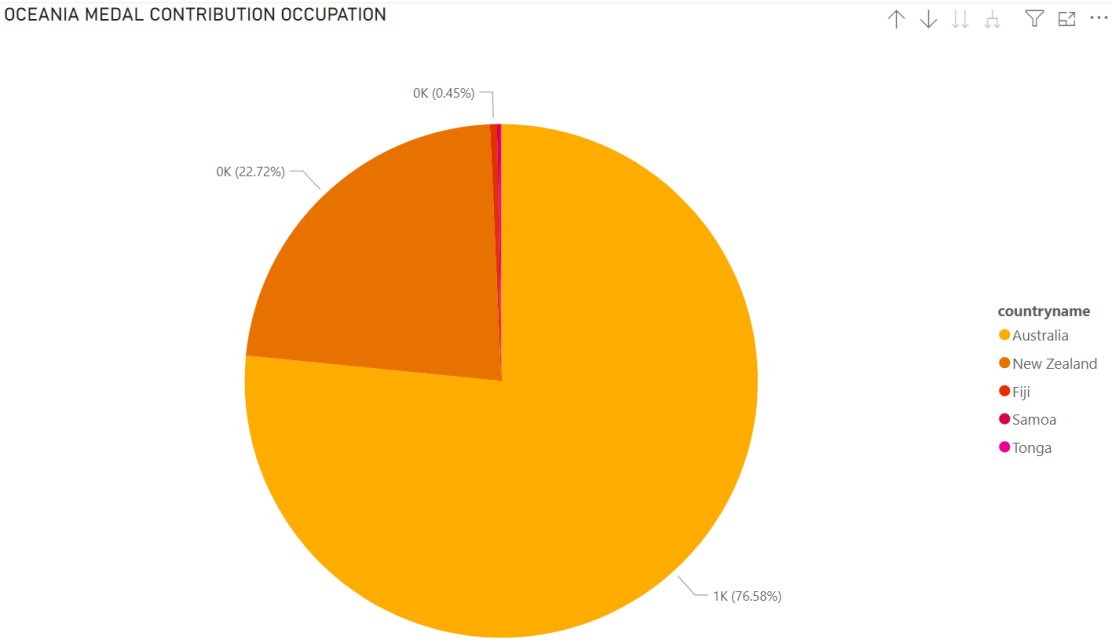


Figure 48: Query 9

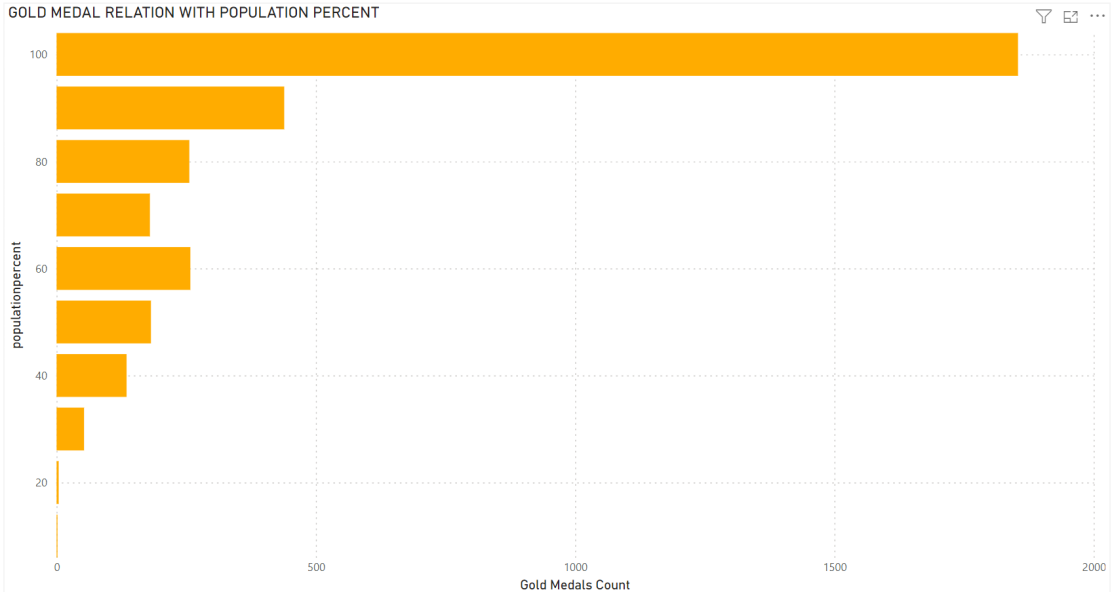


Figure 49: Query 10

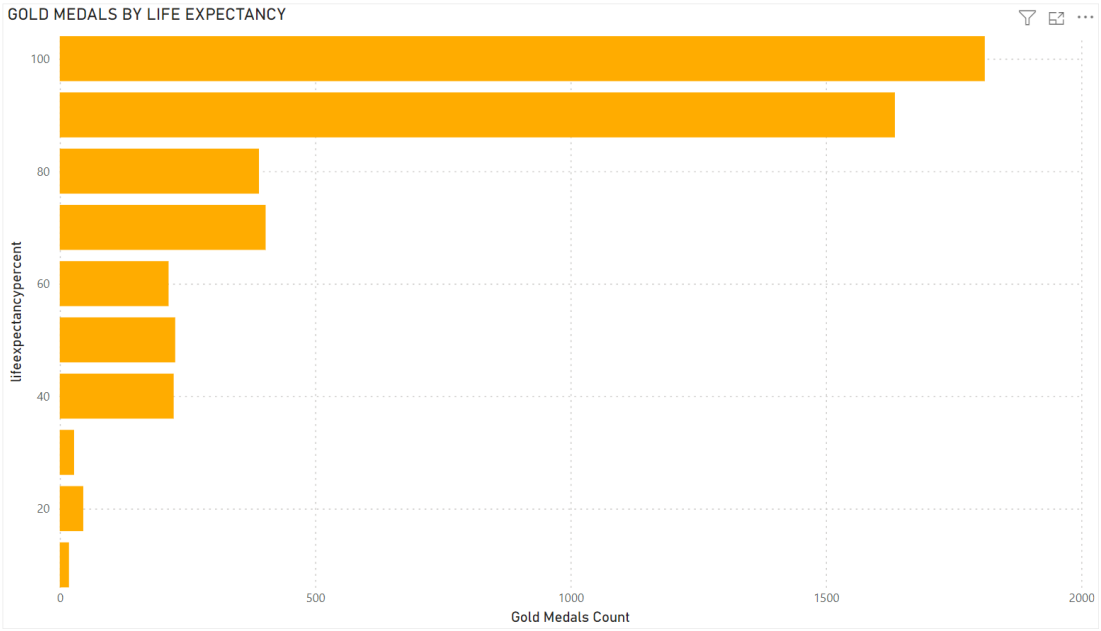


Figure 50: Query 11

3.1.11 Query 11

Is there a correlation between a country's life expectancy and its number of Olympic gold medals?

The data seems better correlated contrast to population. However, this also may be infused by economy and health. For clients need, this graph only focus on the query itself. In conclusion, gold medals are highly correlated to life expectancy. (Fig 50)

3.1.12 Query 12

How do the continents compare in terms of their Olympic medal contributions by decades?

This graph shows that Asian countries significantly grown their influence in Olympics by decades. Africa is also growing but slowly. The IOC may plan to make more efforts on African support in Olympics after collecting the data. (Fig 51)

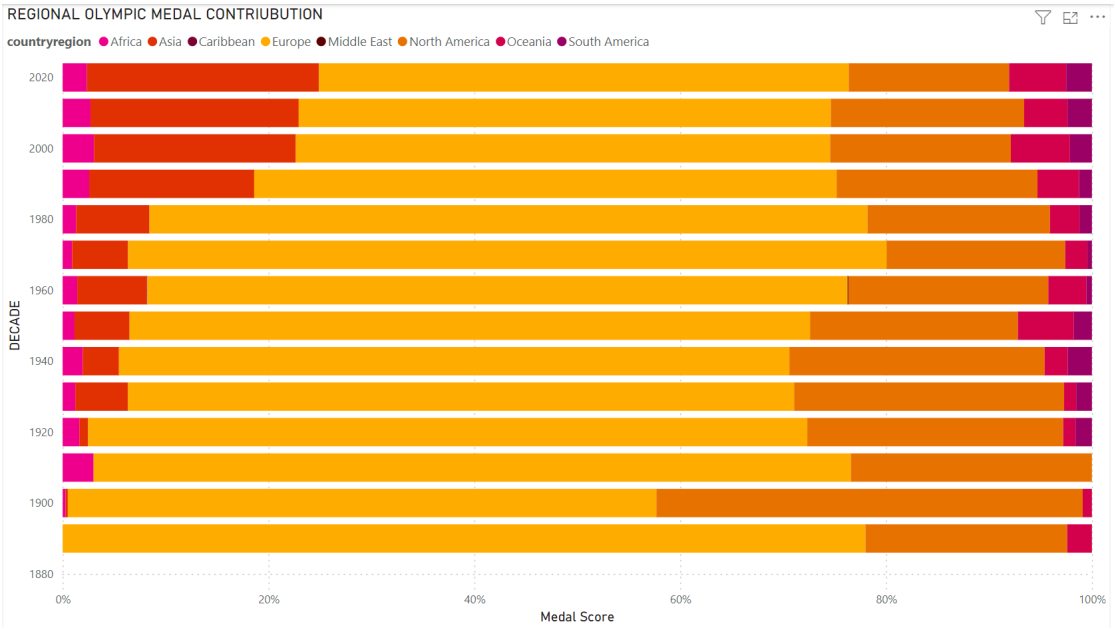


Figure 51: Query 12

3.1.13 Query 13

What are the gold medal statistics for EU countries in the Olympic Games?

This shows an overview data of the EU countries, which might be used in discussing the further questions by making a brief view on it. (Fig 52)

3.1.14 Query 14

How have EU gold medal wins varied by decade in the Olympic Games?

The graph shows that EU countries medals grows by decades but seems to have reached it's peak. EU may need to figure out what caused this and make attempts to fix it. (Fig 53)

3.1.15 Query 15

How does the gender distribution of Olympic medals compare between North America and Europe?

This shows the EU had better gender equality performance than NA, and the EU

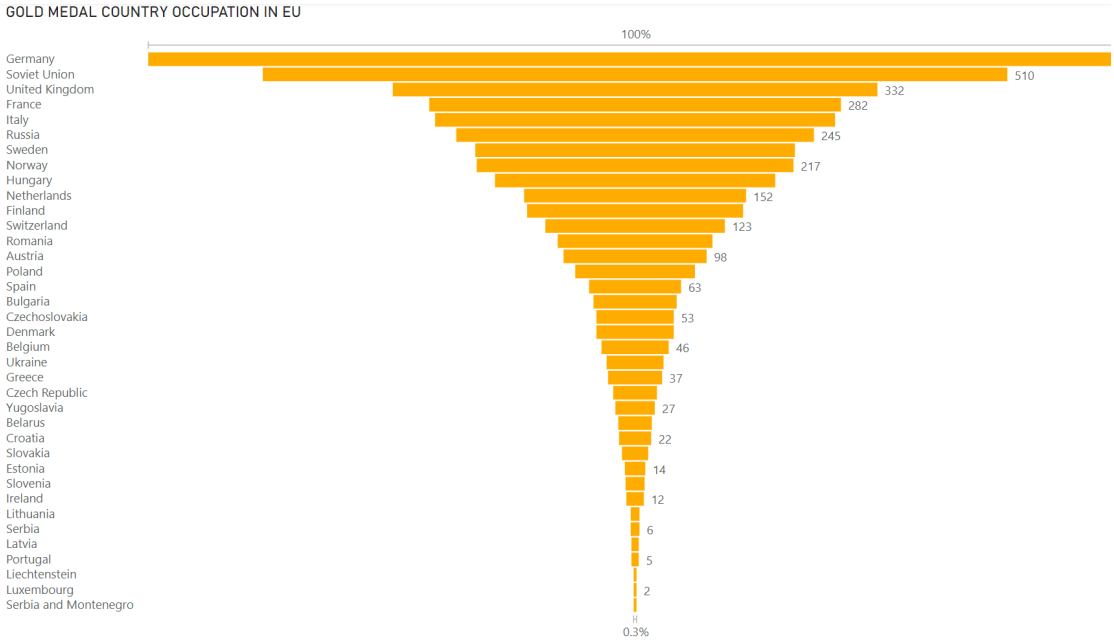


Figure 52: Query 13

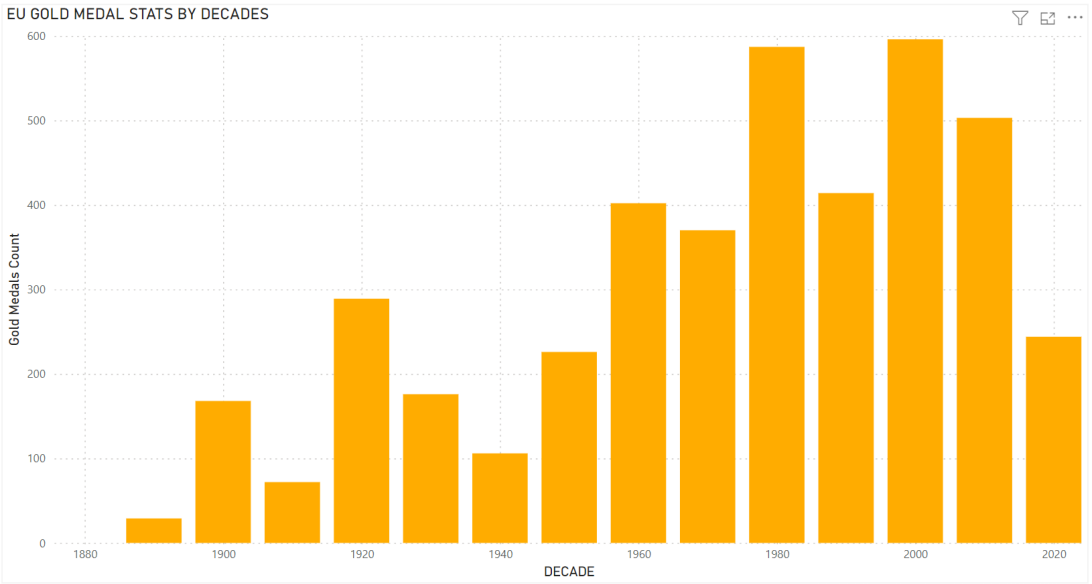


Figure 53: Query 14

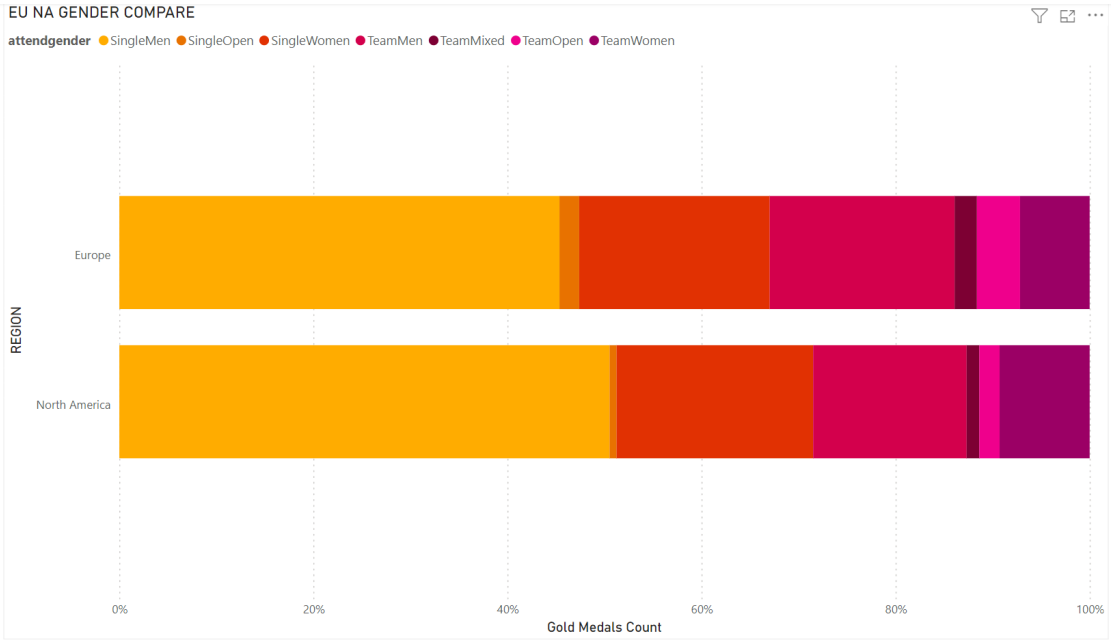


Figure 54: Query 15

could make analysis on the current data to see what they can do for improving the fairness by using the data. (Fig 54)

3.1.16 What If Analysis

The What-if analysis is about if the gdp percent increase 20 percent how regions would change.(Fig 55)

From the data, only medal type could be used for measures. In order to implant the effect of GDP change, a new measure would be created as predicted gold count with a parameter. To modify the gold count by gdp change, a regression model would be created by python.(See PJ1-regression For Whatif Analysis.ipynb)

From the previous graph, it seems following Poisson Model Regression below:

$\log(\mu) = 2.7973 + 0.0193 \times \text{gdppercent}$

After that the measure would be created by DAX, the bar chart could be created to see the change of regions by increse the gdp by 20 percent.

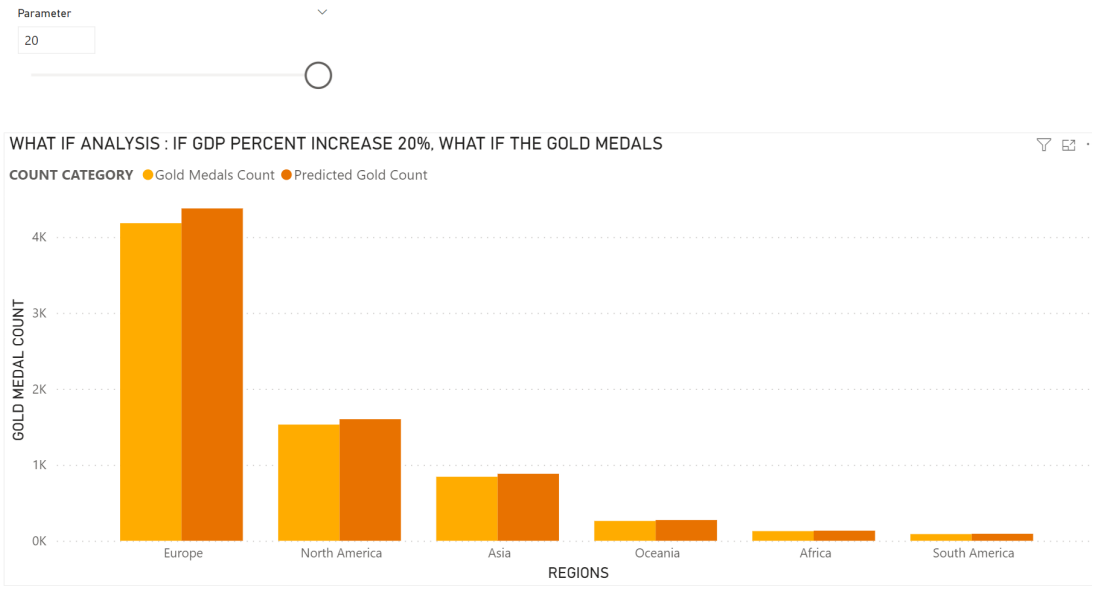


Figure 55: What If Analysis

Due to the source data might be not that precise, the prediction could be more accurate if the data could be the raw data.

4 Association Rule Mining

4.1 Design Explaination

Association Rule Mining is a rule-based machine learning method for discovering interesting relations between variables in large databases. 'mlxtend' will be used library in Python via 'Jupyter lab' on an ipynb file to run the Apriori algorithm. The relationships between the categories will be explored in which countries participate and the medals they receive in each Olympic Games to investigate their correlations.

The meaning of k rules in data mining is finding the top 'k' most common patterns in a dataset, like how items are appeared together, but the rules have to be frequently and valid(beyond the support threshold and confident threshold).[3] This mining process will utilize lift as the criterion for significance.

```
#setting threshold
frequent_itemsets = apriori(transformed_df, min_support=0.04, use_colnames=True, max_len=6)
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.90)
rules = rules.sort_values(by='lift', ascending=False)

print(rules)
```

Figure 56: Threshold

1. ETL, one part of this is contributed by GPT 4
2. Set the disciplines which countries are awarded in games as a transaction, containing multiple disciplines of medals won in that Olympics.
3. Apply the Apriori algorithm for association rule mining. Set the threshold of support to 0.04 to filter the uncommon and confident to 0.90 to withdraw those not strong enough. The result will be listed in the order of lift.(Fig 56)

4.2 Result Analysis

Since our data is sorted by lift, the results indicate that a higher lift value corresponds to a stronger enhancement effect of the lhs combination on rhs. In these rules, their lift values are all greater than or equal to 2, and they all satisfy the principle of k rules. The data reveals that the combination of athletics, diving, and rowing has the most significant enhancement effect on swimming, followed by two rules that also have commendable effects on swimming, and almost all including aquatic sports categories. This demonstrates that cross-training in various categories of aquatic sports has a notable impact. In addition, boxing, diving and athletics also experienced significant lift effects. Boxing, Shooting and weightlifting strongly boost performance in wrestling, while athletics benefits from a very diverse range of enhancements.

From above, several information could be summarized:

1. The performance of athletics in a country may indicates the power of average Olympic level. And it could have a positive effect on other fields.
2. Several disciplines are correlated.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
21	(Rowing, Diving, Athletics)	(Swimming)	0.045636	0.191671	0.042213	0.925000	4.825967	0.033466	10.777714
6	(Rowing, Diving)	(Swimming)	0.047918	0.191671	0.043354	0.904762	4.720380	0.034170	8.487450
28	(Boxing, Shooting, Weightlifting)	(Wrestling)	0.048488	0.247005	0.044495	0.917647	3.715093	0.032518	9.143509
30	(Boxing, Rowing, Swimming, Wrestling)	(Athletics)	0.040502	0.393041	0.040502	1.000000	2.544267	0.024583	inf
0	(Boxing, Diving)	(Athletics)	0.042784	0.393041	0.042213	0.986667	2.510343	0.025398	45.521962
3	(Diving, Swimming)	(Athletics)	0.052481	0.393041	0.051341	0.978261	2.488957	0.030713	27.920137
2	(Diving, Shooting)	(Athletics)	0.045066	0.393041	0.043925	0.974684	2.479855	0.026212	23.974900
20	(Rowing, Diving, Swimming)	(Athletics)	0.043354	0.393041	0.042213	0.973684	2.477313	0.025173	23.064461
26	(Rowing, Swimming, Wrestling)	(Athletics)	0.051911	0.393041	0.050200	0.967033	2.460390	0.029797	18.411105
14	(Boxing, Rowing, Swimming)	(Athletics)	0.056475	0.393041	0.054193	0.959596	2.441468	0.031996	15.022248
25	(Swimming, Rowing, Shooting)	(Athletics)	0.050770	0.393041	0.048488	0.955056	2.429918	0.028534	13.504849
12	(Sailing, Boxing, Rowing)	(Athletics)	0.050200	0.393041	0.047918	0.954545	2.428619	0.028187	13.353109
1	(Rowing, Diving)	(Athletics)	0.047918	0.393041	0.045636	0.952381	2.423111	0.026802	12.746149
4	(Diving, Wrestling)	(Athletics)	0.046777	0.393041	0.044495	0.951220	2.420156	0.026110	12.442670
18	(Canoe Sprint, Shooting, Rowing)	(Athletics)	0.042784	0.393041	0.040502	0.946667	2.408573	0.023686	11.380491
17	(Sailing, Boxing, Swimming)	(Athletics)	0.049629	0.393041	0.046777	0.942529	2.398045	0.027271	10.561095
8	(Boxing, Rowing, Cycling Track)	(Athletics)	0.047347	0.393041	0.044495	0.939759	2.390998	0.025886	10.075528
22	(Rowing, Shooting, Gymnastics Artistic)	(Athletics)	0.047347	0.393041	0.044495	0.939759	2.390998	0.025886	10.075528
27	(Wrestling, Rowing, Weightlifting)	(Athletics)	0.046777	0.393041	0.043925	0.939024	2.389129	0.025539	9.954136
16	(Sailing, Boxing, Shooting)	(Athletics)	0.046207	0.393041	0.043354	0.938272	2.387214	0.025193	9.832744
11	(Boxing, Rowing, Gymnastics Artistic)	(Athletics)	0.045636	0.393041	0.042784	0.937500	2.385250	0.024847	9.711352
10	(Boxing, Fencing, Rowing)	(Athletics)	0.045636	0.393041	0.042784	0.937500	2.385250	0.024847	9.711352
29	(Boxing, Rowing, Shooting, Wrestling)	(Athletics)	0.043925	0.393041	0.041072	0.935065	2.379055	0.023808	9.347176
13	(Boxing, Rowing, Shooting)	(Athletics)	0.054763	0.393041	0.050770	0.927083	2.358748	0.029246	8.324016
23	(Rowing, Gymnastics Artistic, Wrestling)	(Athletics)	0.046777	0.393041	0.043354	0.926829	2.358101	0.024969	8.295113
9	(Boxing, Cycling Track, Swimming)	(Athletics)	0.046207	0.393041	0.042784	0.925926	2.355803	0.024623	8.193953
19	(Canoe Sprint, Swimming, Rowing)	(Athletics)	0.045636	0.393041	0.042213	0.925000	2.353447	0.024277	8.092793
24	(Sailing, Rowing, Wrestling)	(Athletics)	0.043925	0.393041	0.040502	0.922078	2.346012	0.023238	7.789314
15	(Boxing, Rowing, Wrestling)	(Athletics)	0.060468	0.393041	0.055334	0.915094	2.328244	0.031567	7.148634
7	(Boxing, Canoe Sprint, Shooting)	(Athletics)	0.045636	0.393041	0.041643	0.912500	2.321644	0.023706	6.936680
5	(Rowing, Gymnastics Artistic)	(Athletics)	0.063890	0.393041	0.057616	0.901786	2.294384	0.032504	6.179951

Figure 57: Result

3. Some of the correlated disciplines may not be well discovered, like shooting and wrestling.

The commerce advices for clients:

1. For the Australian Olympic Committee (AOC), the insights from the data could guide a strategic push within Australia to combine various aquatic sports for heightened promotional activities. In collaboration with sports facilities and educational insitutes, efforts could be made to encourage venues that offer one type of water sport to accommodate multiple disciplines, thereby fostering an environment conducive to cross-training and athlete recruitment. The AOC and its partner organizations might reap mutual benefits from the enhanced disciplines and potential new bloods of aspiring sports talent.
2. Which should be recommended for IOC are that the IOC engage in targeted advertising of associated sports within individual countries. By promoting a country's strengths and its related disciplines, there is an opportunity to enhance view and audience of Olympic broadcasts.
3. The European Union may consider investing in other sports that are closely linked to their areas of strength. By increasing support for various dominant and associated sports events within Europe, they can aim for a higher return on investment. Additionally, they might explore lesser-known correlations, such as the relation between wrestling and shooting, to boost the activity levels of these sports within the region.

5 Declaring the Dying Data Cube Dead Might Be Premature

In recent years, the application of "data cubes" in OLAP has been a subject of controversy. The article "Database cubes are dead; what is their replacement?" seemingly declared the death of data cubes, highlighting numerous drawbacks associated with them and listing several alternatives. From my perspective, it's

premature to label "data cubes" as an outdated technology, however the era of data warehousing models led by Kimball's dimensional modeling is nearing its end, as advancements in hardware technology have significantly reduced the cost of data analysis, rendering it no longer a challenge.

The reasons people believe "data cubes" are outdated because they are complex, stubborn and costy. They could not support real time analytics as well. and Columnar databases seems to be the most popular databases to replace them. With the improvement in memory capacity and the performance of data storage devices, the problems that "data cubes" were designed to solve in the past had already been solved. In some fields, the traditional ETL process has even given way to ELT[4]. Even those who appreciate the value of data cubes may find their organizations hesitant due to the high costs involved.[5] These signs indicate that the current state of "data cubes" is far from optimistic.

The issues mentioned above, I believe, represent the lamentation of data scientists and engineers over the decline of Kimball's multidimensional data warehouse model. It signifies that the data warehouse design paradigm, as represented by the Kimball approach, has become outdated, although the revised edition attempts to implement data cubes using current industry practices[6], this effort seemingly reflects a sense of powerlessness regarding their obsolescence. Furthermore, the development of ETL tools and hardwares has already simplified the entire process and the consume of memory seems become far less important than before. However, for businesses with mature operations, the established data warehouse models are already sufficient for their needs, proving that data warehouses will continue to exist for a longer period.

Now technologies such as columnar databases and data lakes offer robust solutions to store the data, such as columnar database allows user do processing which must be handled as pre-aggreating. Advanced data processing technology keeps the cost lower, and it validates some of the processing by incresing the effieiciency[7]. In some case, the data pre-aggregation is no longer needed.

In summary, I believe that data cubes still have long lasting life expectancy to stay active in the industry, meeting the business needs of companies in the future.

It would serve as some of the bank management systems which are labeled as legacy. However, the development of AI and storage devices with greater capacity could change everything, posing the real challenge for multidimensional data warehouses. Perhaps in the near future, the industry won't need to process data in the traditional sense, but rather rely on the power of AI to derive meaningful information, or store data in a single dimension without worrying about storage issues.

6 References

[1] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011, pp. 46, 108.

[2] R. Kimball and J. Caserta, *The data warehouse ETL toolkit*. John Wiley & Sons, 2004, pp. 294, 295.

[3] P. Fournier-Viger, C.-W. Wu and V. S. Tseng, 'Mining top-k association rules,' in *Advances in Artificial Intelligence: 25th Canadian Conference on Artificial Intelligence, Canadian AI 2012, Toronto, ON, Canada, May 28-30, 2012. Proceedings 25*, Springer, 2012, pp. 61–73.

[4] S. Jones, 'Are cubes dead?' *SQLServerCentral*, Jun. 2022. [Online]. Available: https://www.sqlservercentral.com/editorials/are-cubes-dead.

[5] C. Chin. 'The rise and fall of the olap cube.' Accessed: 2024-04-09. (Jan. 2020), [Online]. Available: https://www.holistics.io/blog/the-rise-and-fall-of-the-olap-cube/.

[6] J. Han, J. Pei and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2011, p. 11.

[7] D. J. Abadi, S. Madden and N. Hachem, 'Column-stores vs. row-stores: How different are they really?' In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ACM, 2008, pp. 967–980.