Project 1: Olympic Athletes Analysis

Client and Business Objectives

Our client aims to derive actionable insights from historical Olympic data, focusing on medal achievements related to athletes' countries and sporting events. Identified clients are important for several reasons:

- By understanding past performance trends, clients can identify strengths and weaknesses across different sports which can help guide strategic decision making on training investments and focus areas.
- Analysing the relationship between a country's economic indicators like GDP and health expenditure with Olympic success can help in lobbying for or justifying increased sports funding from government or private sectors.
- Insights from the data can influence sports policy development, focusing on enhancing facilities and support systems that have historically led to higher medal counts.

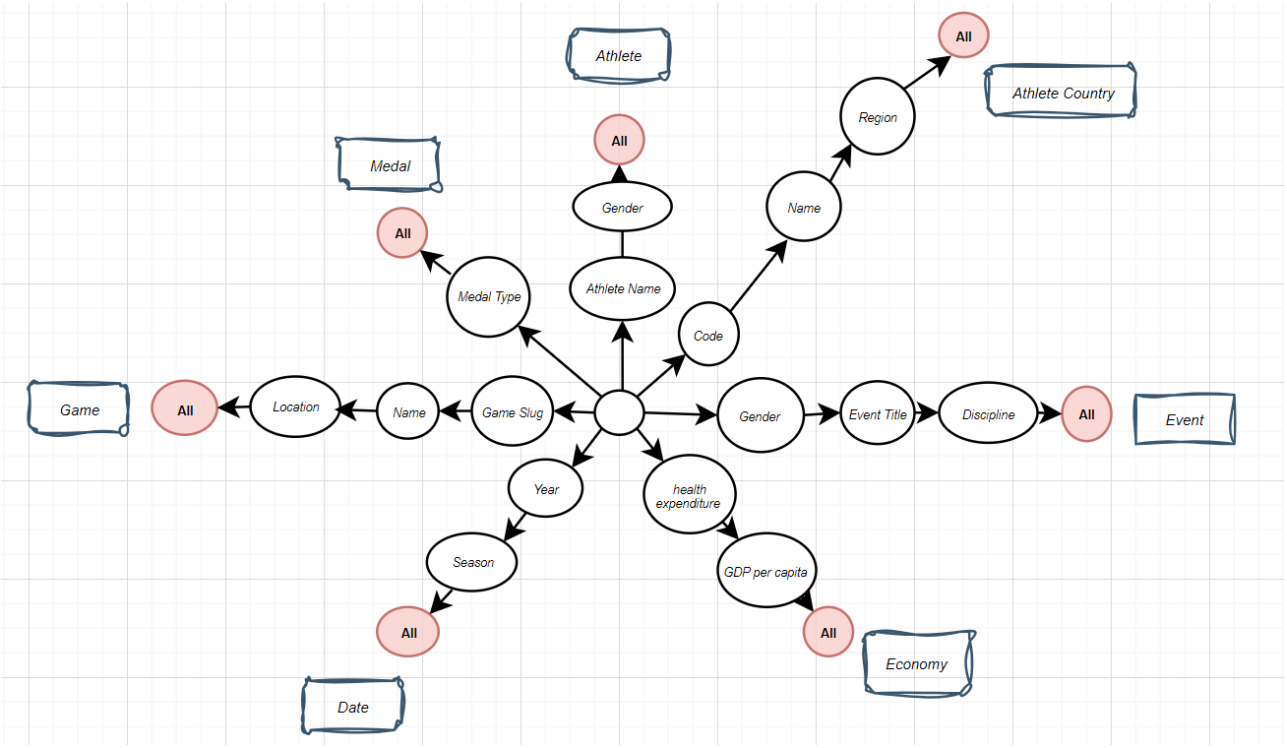
Business Queries

1. **Olympic Performance Analysis:** Which Olympic games, categorized by year and sport, have accumulated the most medals? This analysis further delineates results by medal type.
2. **Economic Impact on Performance:** How does a country's GDP correlate with its Olympic medal tally? This query is explored with breakdowns by medal type to provide nuanced economic insights.
3. **Athlete Participation Trends:** Which Olympic games have seen the highest number of athlete participations? This helps in understanding the scale and engagement levels of different Olympic events.
4. **Decadal Trends by Continent:** What has been the trend in medal counts over the past decade within the specific continent such as Asia?
5. **Seasonal Impact Analysis:** How does seasonality affect medal counts across different continents, such as Asia?

Data Warehouse Design and Concept Hierarchies

Finest Grain

The finest granularity of our data warehouse captures each athlete's participation in specific events, their country of representation, the particular games attended, and the specific medals won.



Athlete Dimension

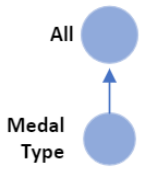
Athlete Dimension stores data on athlete names and gender. It allows for more fine-grained analysis on the effect of event gender or athlete gender on medal count and participation rate from various countries. Athlete Dimension has two levels: Athlete Name, and Athlete Gender. Athlete Genders were derived from event gender which included: 'Men', 'Women', 'Mixed' and 'Open'. This dimension allows for analysis of performance trends across genders and the impact of athlete demographics on Olympic outcomes. It supports gender-based performance comparisons within and across sports disciplines.

Athlete Country Dimension

Athlete Country Dimension stores data on athletes' country. More specifically, athlete country code, country, and continent. It provides critical insights into how the socioeconomic status and geographical location of an athlete's country impact Olympic success. This dimension is crucial for evaluating how well countries with different economic levels perform in terms of medal counts, which can influence national sports funding and development policies.

Medal Dimension

Medal Dimension stores information on medal type. Separating medals by type allows analysts to drill down into the data to understand specific outcomes of the competitions. This dimension is essential for recognizing trends in medal distribution, such as identifying which countries consistently achieve high medal counts and which sports yield specific types of medals.



Event Dimension

Event Dimension stores information on which event the athlete participated in. More specifically, it stores information the event gender, event title, and the sport type / discipline. This dimension offers a granular view of athlete engagement and success rates across different sports disciplines. It supports strategic decisions such as identifying sports and/or events that may require more funding or development based on their popularity and success rate in past Olympics.

Game Dimension

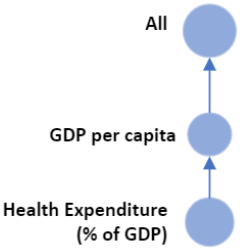
Game Dimension stores information on Olympic Games the athletes were participating in. More specifically, it stores information on the name of the Olympic Game and location i.e., the country the Game was being held in. Analysing games by their geographic and temporal context helps to understand how different locations and their unique characteristics can potentially influence the games' outcomes and athletes' performances.

Year Dimension

Year Dimension stores information on the time period of the Olympic Games in which the athletes were participating in. More specifically, it stores information on the year, and season of the games. This dimension is particularly valuable for examining how external factors such as global economic conditions or other factors such as technological advances in sports training impact Olympic games outcomes and medal counts over different years.

Economy Dimension

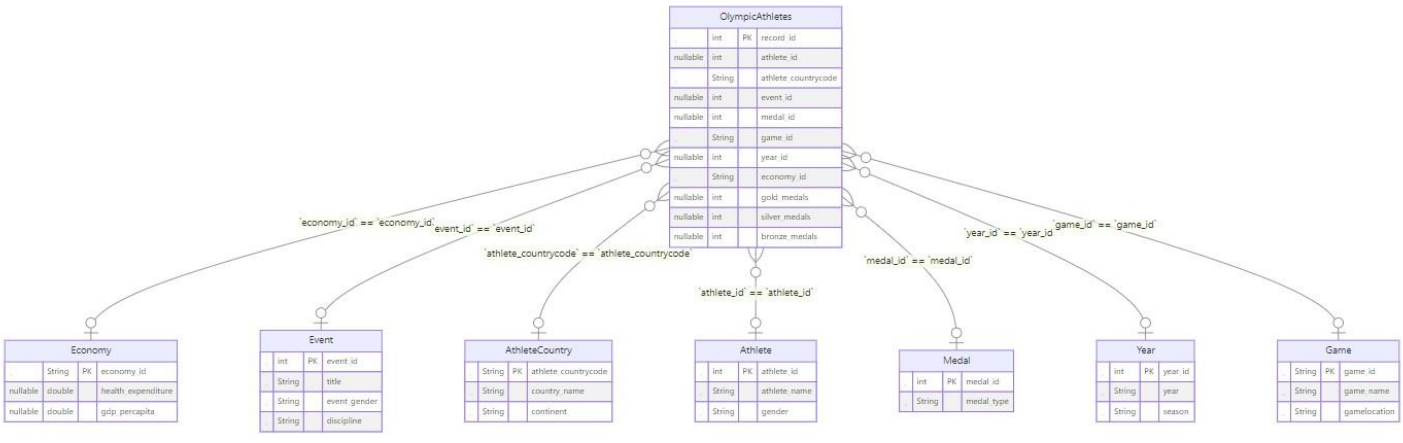
Economy Dimension stores information on economic metrics of the athletes' countries. It includes GDP per capita and health expenditure (% of GDP) in $US. This dimension ties athletes' performance to their country's economic conditions, providing a basis for analysing how economic factors like GDP and health expenditures relate to sports success. This analysis can guide decisions on sports investments and highlight the need for economic support to improve Olympic outcomes.

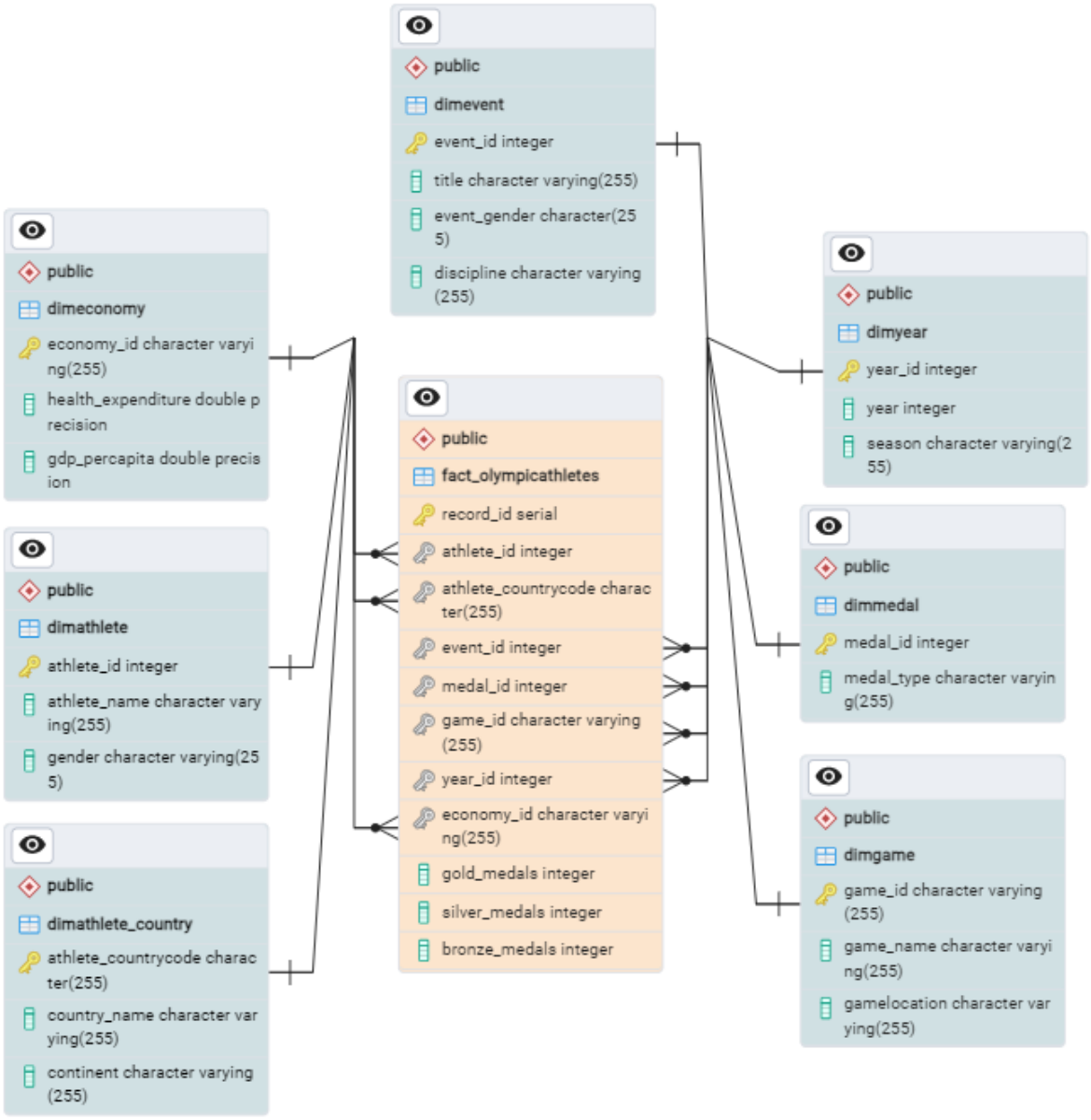


Fact Table

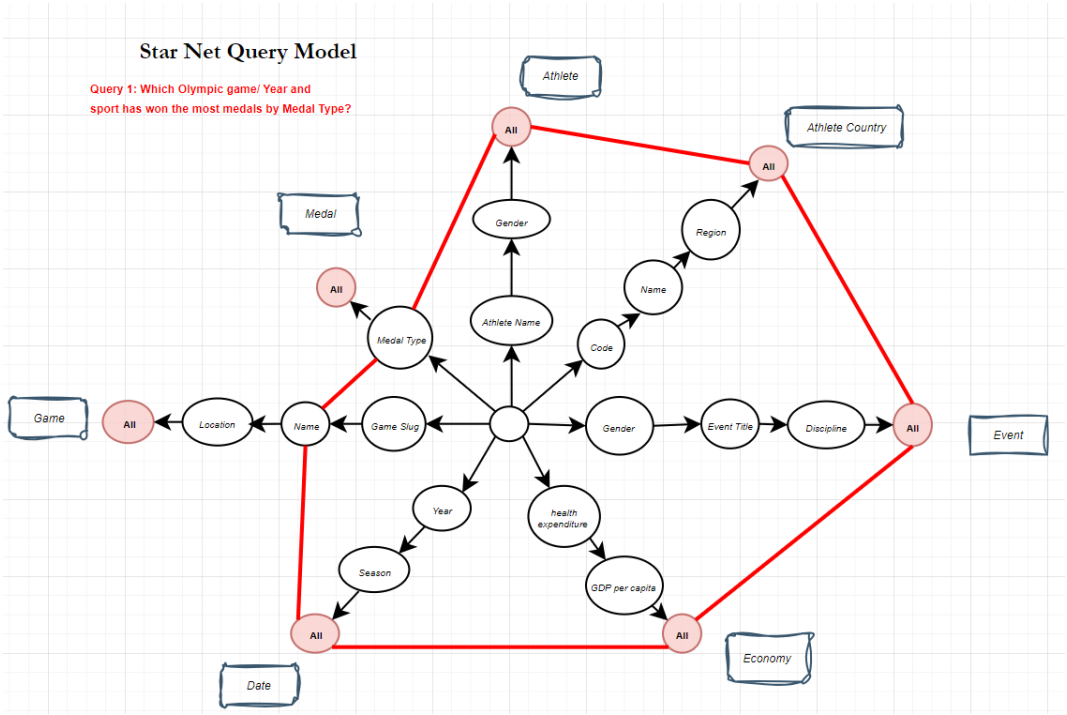
Central to the design, fact table captures all event participations, linking athletes, the medals won, and the corresponding Olympic games. It records each athlete's participation, noting the specific event, the country represented, the game attended, and the medals won. This table is structured primarily with foreign keys linking to dimension tables and includes Boolean fields (1s and 0s) for medal types—gold, silver, and bronze—to facilitate straightforward aggregation of medal counts. While it primarily tracks event participations and medal achievements, its design supports extensive analytical queries through the integration with multiple dimensions, enabling sophisticated roll-up and drill-down analyses within our data warehouse. This setup ensures comprehensive insights into the performance patterns and outcomes of Olympic participants. ERD diagram for database and query footprints are as follow

ERD diagram and query footprints:

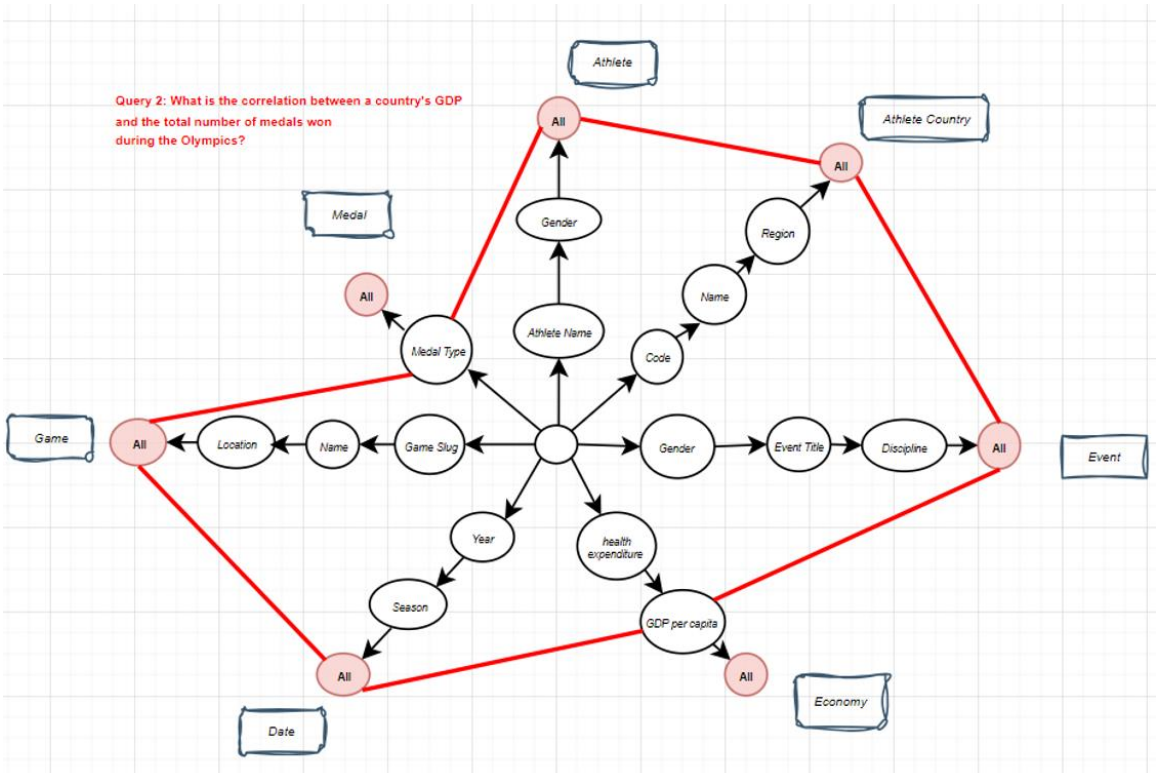

OlympicAthletes
record_id
athlete_id
athlete_countrycode
event_id
medal_id
game_id
year_id
economy_id
gold_medals
silver_medals
bronze_medals
'economy_id' == 'economy_id'
'event_id' == 'event_id'
'athlete_countrycode' == 'athlete_countrycode'
'athlete_id' == 'athlete_id'
'medal_id' == 'medal_id'
'year_id' == 'year_id'
'game_id' == 'game_id'
Economy
economy_id
health_expenditure
gdp_percapita
Event
event_id
title
event_gender
discipline
AthleteCountry
athlete_countrycode
country_name
continent
Athlete
athlete_id
athlete_name
gender
Medal
medal_id
medal_type
Year
year_id
year
season
Game
game_id
game_name
gamelocation


public
dimevent
event_id integer
title character varying(255)
event_gender character(255)
discipline character varying(255)
public
dimeconomy
economy_id character varying(255)
health_expenditure double precision
gdp_percapita double precision
public
dimyear
year_id integer
year integer
season character varying(255)
public
fact_olympicathletes
record_id serial
athlete_id integer
athlete_countrycode character(255)
event_id integer
medal_id integer
game_id character varying(255)
year_id integer
economy_id character varying(255)
gold_medals integer
silver_medals integer
bronze_medals integer
public
dimmedal
medal_id integer
medal_type character varying(255)
public
dimathlete
athlete_id integer
athlete_name character varying(255)
gender character varying(255)
public
dimgame
game_id character varying(255)
game_name character varying(255)
gamelocation character varying(255)
public
dimathlete_country
athlete_countrycode character(255)
country_name character varying(255)
continent character varying(255)

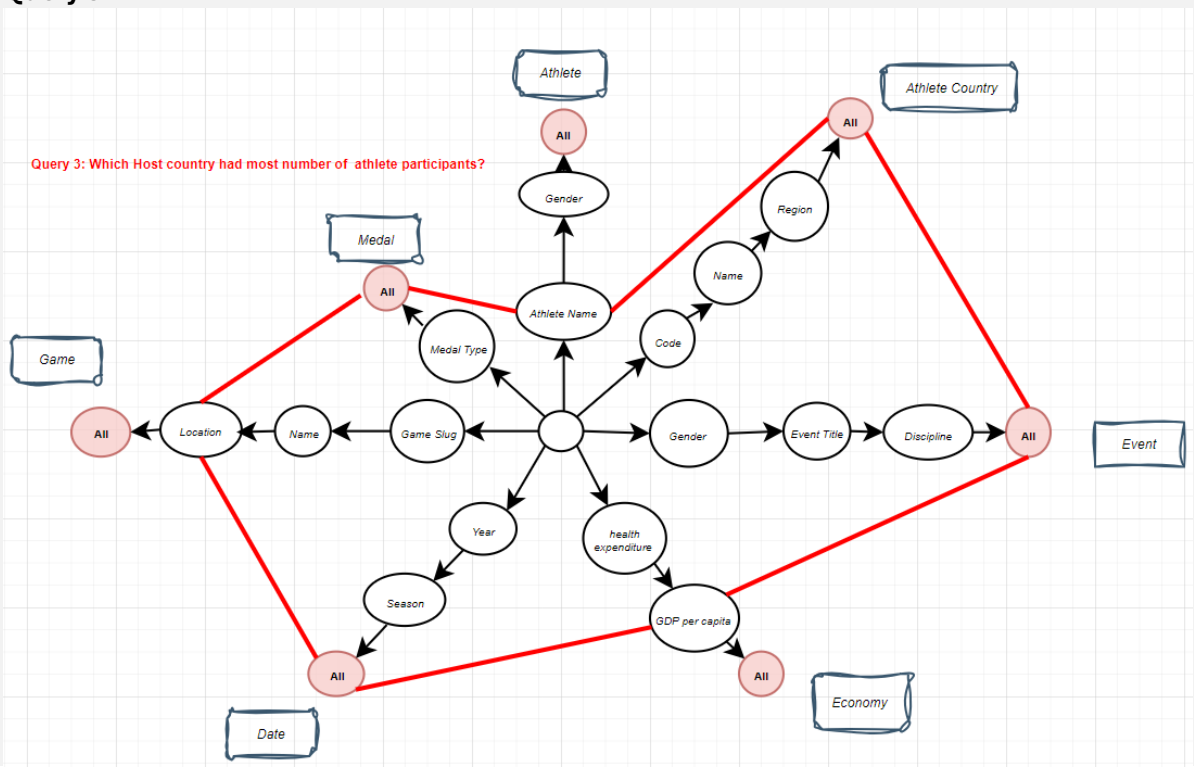
Query 1:



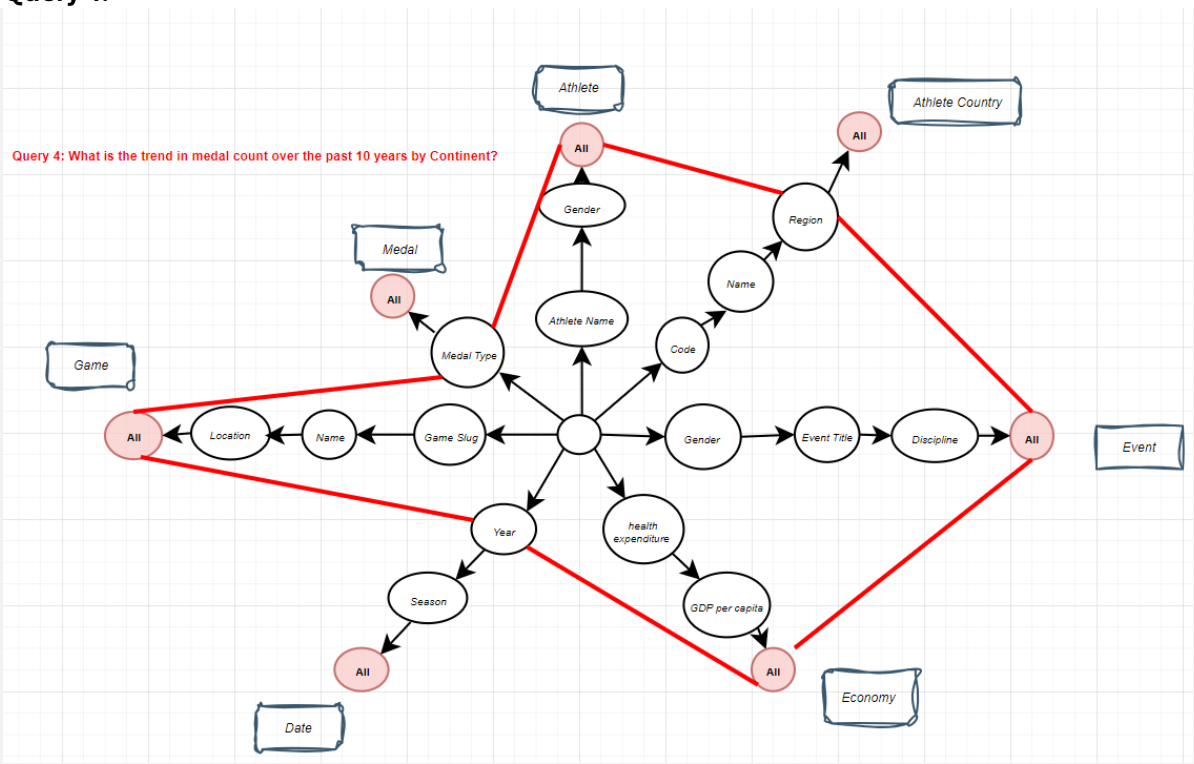
Query 2:



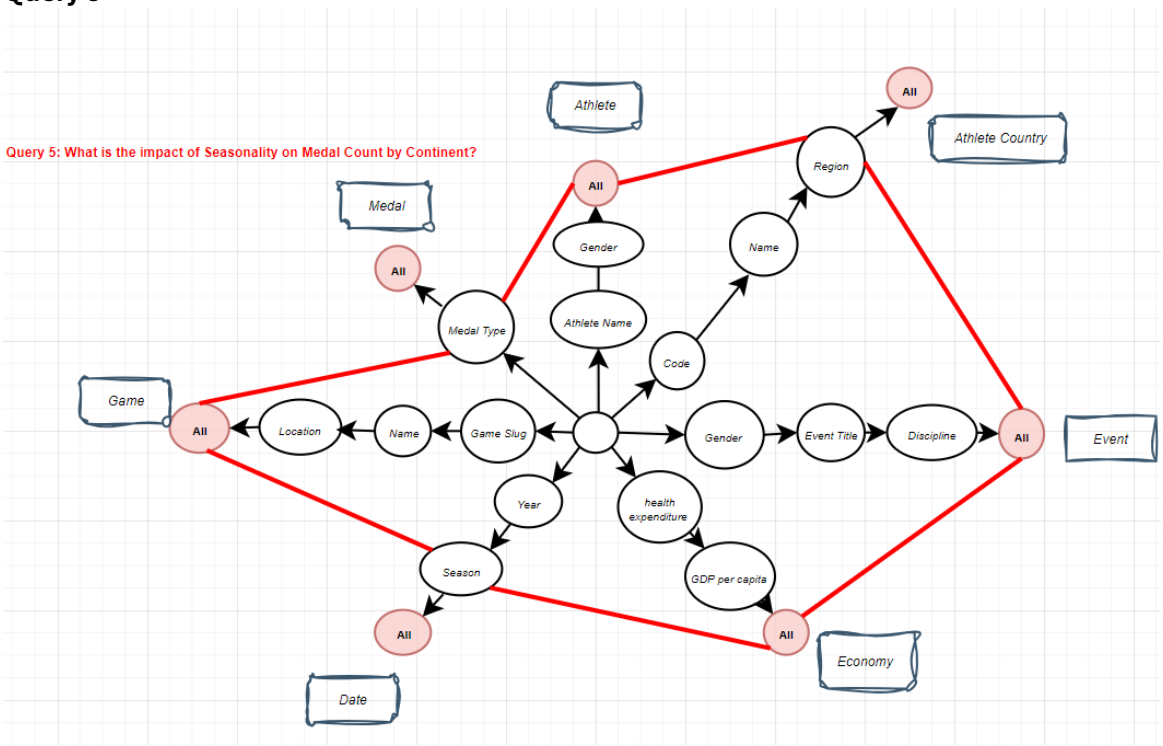
Query 3:



Query 4:



Query 5



Extract, Transform and Load (ETL) Process:

Extract Data

The extraction phase of the ETL process in this project involved collecting data from various prescribed datasets related to Olympic events and associated socio-economic factors. These datasets include historical records of Olympic hosts and medal winners, as well as broader data on mental health, life expectancy, economic indicators, and demographic breakdowns by continent. The datasets were acquired from reputable organizations like the International Monetary Fund (IMF), Our World in Data, and World Population Review, and included structured CSV files detailing Olympic Games history from 1896 to 2022.

Initial data inspection was performed in Excel where irrelevant rows were omitted for streamlined analysis. For example, entries representing collective entities like the USSR or various economic groupings were excluded because they do not correspond to single countries, which are our primary focus for analysis regarding athletes' performances and national participation. Also, redundant columns such as 'Time' and 'Time Code' from the economic data were reduced to single, relevant entries to eliminate redundancy.

Additionally, the Olympic medals dataset was refined by dropping columns with excessive missing data or those less relevant to our analysis goals, such as, 'athlete_url'. This column was sparsely filled and was not the focus of the analysis. This pruning helped to focus the datasets on the most useful and reliable data. setting the stage for the transformation phase where these curated datasets would be further processed and integrated into our data warehouse. After dropping un-necessary columns and rows data was loaded in python was further transformation.

Transform Data

The pre-processed data was loaded into a Python environment using Jupyter Notebook, with Pandas being the primary tool for data cleaning and transformation. The transformation process involved several steps, such as standardizing country codes and names, unifying column names, adding identifier columns, removing duplicates, and handling missing values by either imputing new values or replacing them with zeros. These steps were crucial in preparing the data for further analysis and integration into the data warehouse. Notably, data regarding life expectancy and mental illness were excluded from the analysis. This decision was made because these datasets did not directly relate to the project's focus on analysing Olympic performance through economic factors and geographical origins. Including such health data would have required complex modelling to discern any meaningful insights, potentially introducing variables that could obscure the analysis's primary objectives.

Database Creation in PostgreSQL: Before data transformation, data base was designed and created in PostgreSQL and python notebook was connected to the database.

Data transformation process is as follows:

1. **Clean country region table:** Country region table initially consisted of two columns: Country and Region. There were a lot of Athletes countries that were not included in the country region table. The life-expectancy table, containing accurate country names and codes along with extensive country data, facilitated the merging of relevant tables. This integration was essential for aligning demographic information with other datasets. This resulted in original country region table having a code table and entities that were not initially present. Final merged table was checked for any NANs and values were manually updated with correct country code and country names. Rows with country names that were duplicated with two different names were removed to only keep one.
2. **Clean Medals Data:** The initial step involved identifying and addressing missing data, particularly in Athlete Names. A unique identifier was constructed for each athlete by concatenating 'Medal type,' 'event title,' and 'country code' to compensate for missing names. Subsequently, two letter 'country code' and 'participant type' columns were removed from the analysis due to sparse data and irrelevance to the study's focus. Additionally, the two-letter country codes were discarded in favour of three-letter codes to align with analysis requirements.
3. **Clean Economic Data:** Economic data was refined by selecting key economic indicators such as GDP, health expenditure, and corresponding country codes. Additionally, columns were renamed for clarity and ease of access, ensuring that the names are descriptive and facilitate straightforward data analysis.
4. **Modify Medals Data:** Since medals data consisted of most columns necessary for analysis it was treated as a subset of a fact table. Athlete ID, Event ID, Game ID and Medal ID columns were added to the medal data, and columns were renamed for consistency.
5. **Populate Medal ID:** Then Medal ID dictionary was created to map Medal type and Medal ID into the medals data and populate Medal ID columns. Medal Dimension was separately created due to its simplicity.
6. **Correct Country Names and Country Codes for consistency:** Medals Data and Countries by Region data was examined in detail, and it was revealed that few countries names and country codes were consistent. Since country codes will be used as primary key for Athlete Country Dimension, country names and country codes were corrected for medals and country by region data.
7. **Athlete Country Dimension:** Using Medals Data athlete country dimension was separated which initially included athlete ID (to only include countries of Athletes), Athlete Full Name, Country code and country Name. Any duplicates were removed. There were few entities

such as ‘USSR’ and ‘MIX’ that few athletes were representing, and regions was missing in the data. Hence, the region was mapped and included in the country region column.

8. **Athlete Dimension:** Athlete Dimension was created by separating athlete attributes such as athlete ID, athlete full name, event gender, medal type and medal ID. To populate Athlete ID, athlete dimension index was used. To create a gender column for the athletes, using event gender, ‘Event Gender Prioritization’ was used. i.e., to ensure that 'Men' or 'Women' entries are prioritized over 'Mixed' or 'Open', a priority dictionary was created to map each gender type to a numerical value, where 'Men' and 'Women' have a higher priority. The data frame was sorted based on athlete_full_name and gender_priority. This sorting ensured that entries where athletes are identified distinctly as 'Men' or 'Women' come before entries classified as 'Mixed' or 'Open'. After sorting, duplicate entries for athletes (based on athlete_full_name) were removed, keeping only the first entry which, due to prior sorting, should correctly represent their gender if specified as 'Men' or 'Women'. The resulting data frame is then reduced to just the columns Athlete_ID, athlete_full_name, and event_gender, and the columns are renamed to 'Athlete Name' and 'Gender' to reflect their contents more accurately.
9. **Athlete Country Dimension Cleaning:** Athlete Country Dimension from Step 7 and Athlete Dimension from Step 8 were joined together to populate athlete ID in athlete country dimension. This step was necessary to only keep rows or country data of athletes. The joined dataset was examined for athletes who are listed under multiple countries. Instances where an athlete is associated with more than one country were filtered out, specifically aiming to resolve cases where an athlete is listed under both a specific country and an 'International' designation. To prioritize athletes' specific country representation over 'International' or less specific entries, data was sorted. After sorting, duplicate entries were removed for each athlete, keeping only the entry for the specific country of representation, and excluding 'International' unless it's the only available option. Data frame was cleaned by dropping any unnecessary or temporary columns used during the sorting and filtering processes to ensure the dataset reflects only unique, accurate representations for each athlete's country. Any duplicated rows were dropped to ensure all rows are unique. Columns were renamed for clarity and consistency as required, and ensure the final dataset is devoid of any remaining duplicates. This dataset can then be utilized for deeper analysis related to athletes' performance and demographic trends.
10. **Populate Athlete ID:** Finally, athlete Dimension was merged with medal data to populate correct Athlete IDs.
11. **Event Dimension:** Event ID was initially created in medals data by concentrating ‘discipline title’, ‘slug game’, ‘event title’ and ‘event gender’. The resulting strings were the factorised to create an ‘event identifier’ and added to new column ‘Event ID’. Finally, all unnecessary columns such as ‘event identifier’ was dropped and only Event ID was kept. Event Dimension table was created that included: Event ID, event gender, event title, and discipline title. Any duplicates were dropped from the dimension table to ensure all rows were unique.
12. **Economic Dimension:** Country Codes from Economic subset data frame from Step 3 were standardised to ensure consistent country codes before merging the data with medals data. Economic ID were created using country codes since there were only unique country codes in economic dimensions. Finally economic ID was populated in medals data and fact and medals data were separated.
13. **Game Dimension:** The Game Dimension was constructed from host country data focusing on 'game slug', 'game name', and 'game location'. The 'game slug' was designated as the game ID due to its uniqueness across the dataset. To ensure data integrity, all columns were

converted to appropriate data types and duplicates were removed. This dimension was kept separate in the data model, using 'game slug' as a link to the fact table without needing to merge, as the fact table already contained matching game slugs.

14. **Year Dimension:** Similar approach was used as Game Dimension to create Year Dimension. Hosts data was used to create Year Dimension and include columns such as 'game id', 'game season' and game year. Any duplicates were removed to keep the rows unique and data types were converted to relevant types. Year ID was created by setting the index of the data frame as ID. Finally, year dimension was joined to fact table.
15. **Clean Fact Table:** Fact table was cleaned to include only ID columns and any duplicated rows were removed.
16. **Final check up and cleaning:** It was later found that health expenditure column from economic dimension has values '..' which were not relevant hence they were replaced with 0. All dimension tables were checked to ensure they only have relevant columns. All the column names for all fact and dimension tables were renamed to ensure they match the column names in PostgreSQL database.

Load Data

Finally, all the dimension and fact table were loaded into PostgreSQL using psyopg2 library. After loading data into the database measure columns were added to the fact table which included: count of gold medal, count of silver medals and count of bronze medals. These measure columns allowed for easier aggregation of medal wins by medal type and contained values of 0s and 1s due to nature of graduality of fact row i.e., one athlete can only win one medal per event.

Multi-dimensional Analysis:

For each dimension hierarchies were created using Atoti multi-dimensional cubes to conduct sophisticated roll-up and drill-down operations. Hierarchies were created based on Star Net diagram. Since GDP and Health Expenditures were numerical values, it was hard to implement hierarchies in Atoti however, multi-dimensional analysis was still applied to them. Following explain how cubes can be utilised for multi-dimensional analysis:

1. **Roll-Up**: This operation aggregated data from lower to higher levels of data granularity within the hierarchies. For example, you could start with the total number of medals by athlete, then roll-up to the total medals per sport, per country, or per continent.
2. **Drill-Down**: Opposite to roll-up, this involves breaking down data from higher to lower levels of hierarchy. You might examine overall medal counts by continent and then drill down to specific countries within Asia to see finer details.
3. **Slicing and Dicing**: With Atoti, you can 'slice' to filter data on one dimension (like only looking at the Winter Games) and 'dice' to analyse a subset of data across multiple dimensions (like comparing medal counts across different sports within the Winter Games).
4. **Pivoting**: By pivoting, you can reorient the cube to view it from different perspectives. For instance, pivot by year to see how medal distribution changes over time or by GDP to understand economic impacts.

For example, in the following query the cube function aggregates the sum of gold medals across different disciplines and splits them by gender (event gender). This showcases a roll-up operation where individual medal counts are aggregated to a higher level within the hierarchy (discipline), which can be broken down further into sub-categories (gender). This capability allows for complex, nuanced analysis, enabling the user to drill down into specific areas of interest (e.g., looking at gold medals in Judo for Men) or to roll-up to see larger trends (e.g., total gold medals across all disciplines).

```
cube.query(measures["gold_medals.SUM"], levels=[levels["discipline"]])
```

✓ 0.3s

			gold_medals.SUM
event_gender	title	discipline	
Men	+ 100kg (heavyweight) men	Judo	2
	+ 80 kg men	Taekwondo	1
	+ 91kg (super heavyweight) men	Boxing	2
	+ 95kg (heavyweight) men	Judo	1
	+105kg men	Weightlifting	1
...	...	...	...
Women	vault women	Gymnastics Artistic	18
	volleyball women	Volleyball	14
	water polo women	Water Polo	5
	épée individual women	Fencing	6
	épée team women	Fencing	5

Similarly, by querying the total medals by continent, you're able to perform a roll-up analysis to see the total medal count for each continent, which is a higher level of aggregation. By having country and continent as levels within the dimension, you can drill down to see the medal count for individual countries or roll up to view the aggregate for continents. This assists in understanding the global distribution of athletic success and could inform decisions on international sports development strategies.

```
cube.query(measures["Total Medals.SUM"], levels=[levels["continent"]])
```

✓ 0.3s

		Total Medals.SUM
country_name	continent	
Afghanistan	Asia	2
Algeria	Africa	17
Argentina	South America	87
Armenia	Asia	18
Australasia	Oceania	12
...	...	...
Vietnam	Asia	4
Virgin Islands (U.S.)	North America	1
West Indies Federation	North America/Caribbean	2
Zambia	Africa	2
Zimbabwe	Africa	8

Furthermore, Business Queries were queries using SQL.

SQL Queries Output:

Query 1:

	game_name character varying (255)	total_gold_medals bigint	total_silver_medals bigint	total_bronze_medals bigint
1	[null]	7109	7059	7529
2	Tokyo 2020	376	374	438
3	Rio 2016	337	337	389
4	Athens 2004	333	332	357
5	Beijing 2008	332	332	383
6	Sydney 2000	332	332	359
7	London 2012	331	333	380
8	Atlanta 1996	296	298	323
9	Barcelona 1992	282	279	326
10	Seoul 1988	260	252	285

Query 2:

	country_name character varying (255)	gold_medals bigint	silver_medals bigint	bronze_medals bigint	gdp_percapita double precision
1	United States	1222	1000	872	63528.6343
2	Germany	667	667	661	46772.82535
3	China	333	260	214	10408.71912
4	United Kingdom	332	363	350	40318.41692
5	France	282	305	365	39055.28293
6	Italy	274	243	288	31918.69349
7	Russia	245	236	255	10194.44141
8	Sweden	219	241	249	52837.90398
9	Norway	217	192	180	68340.0181
10	Hungary	192	166	201	16125.60941

Query 4:

	year integer	total_medals bigint
1	2004	216
2	2006	25
3	2008	253
4	2010	33
5	2012	258
6	2014	26
7	2016	264
8	2018	41
9	2020	302
10	2022	43
11	[null]	1461

Query 3:

	host_country character varying (255)	number_of_athletes bigint
1	United States	2428
2	Japan	1886
3	Great Britain	1689
4	China	1276
5	Greece	1042
6	Republic of Korea	1011
7	Brazil	997
8	Australia	953
9	Canada	947
10	France	864

Query 5:

	season character varying (255)	year integer	total_medals bigint
1	Summer	2004	216
2	Winter	2006	25
3	Summer	2008	253
4	Winter	2010	33
5	Summer	2012	258
6	Winter	2014	26
7	Summer	2016	264
8	Winter	2018	41
9	Summer	2020	302
10	Winter	2022	43
11	Summer	[null]	1293
12	Winter	[null]	168
13	[null]	[null]	1461

Power BI Data Visualization:

Power BI can significantly enhance the visualization of query results from the Olympics database by leveraging its capacity for multi-dimensional analysis. With its intuitive interface, Power BI allows users to easily create dashboards that enable both roll-up and drill-down functionalities. This means you can aggregate data to see overall trends, like total Olympic medals won by continent, and then seamlessly drill down to examine more specific data points, like the medal distribution for individual countries within those continents.

For example, Power BI can display hierarchies such as the athlete's country, the events they participated in, and the medals won, allowing for a detailed yet comprehensive visual analysis. You can explore data across different dimensions like time (year dimension) or location (game dimension) with interactive charts and graphs that update dynamically, giving stakeholders the ability to spot trends, patterns, and outliers effectively.

By employing Power BI for the Olympic project, one could interactively analyse the historical performance of athletes, compare medal tallies across different sports or Olympic games, and evaluate the impact of economic factors on countries' Olympic success, all within a user-friendly and visually compelling environment.

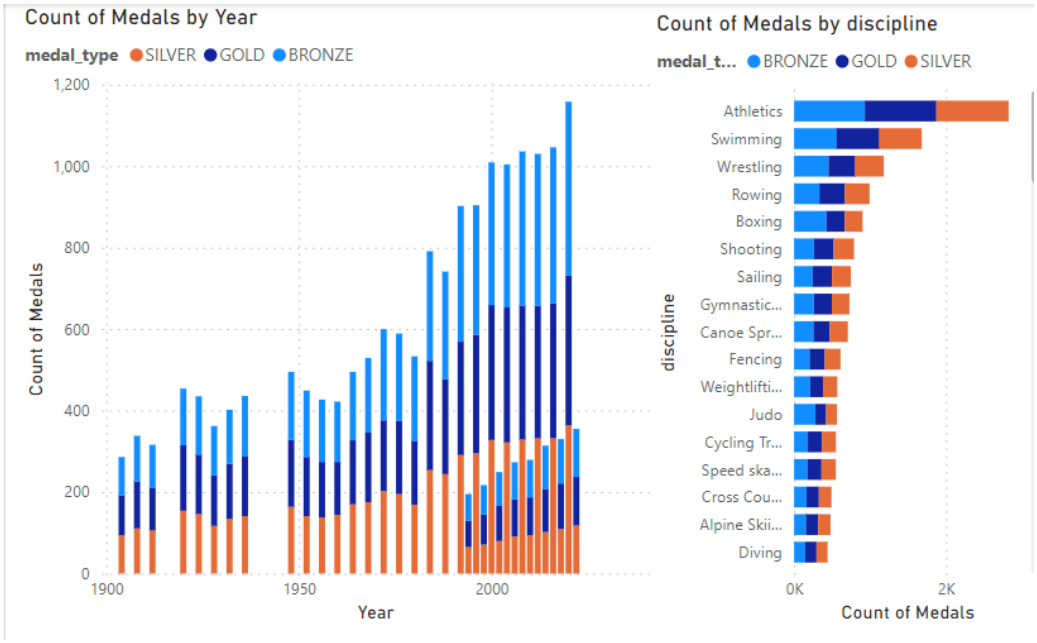
The provided visualizations, SQL queries, and data cube queries facilitate answers to several business inquiries regarding the Olympics:

1. **Medal Accumulation Trends:** An analysis of the Olympic Games, sorted by year and sport, reveals that the total number of medals won has increased over time. This trend might be attributed to growing global populations and the expansion of training and support facilities.

Bronze medals are the most frequently awarded, with athletics, swimming, and wrestling being the top disciplines for medal winners.

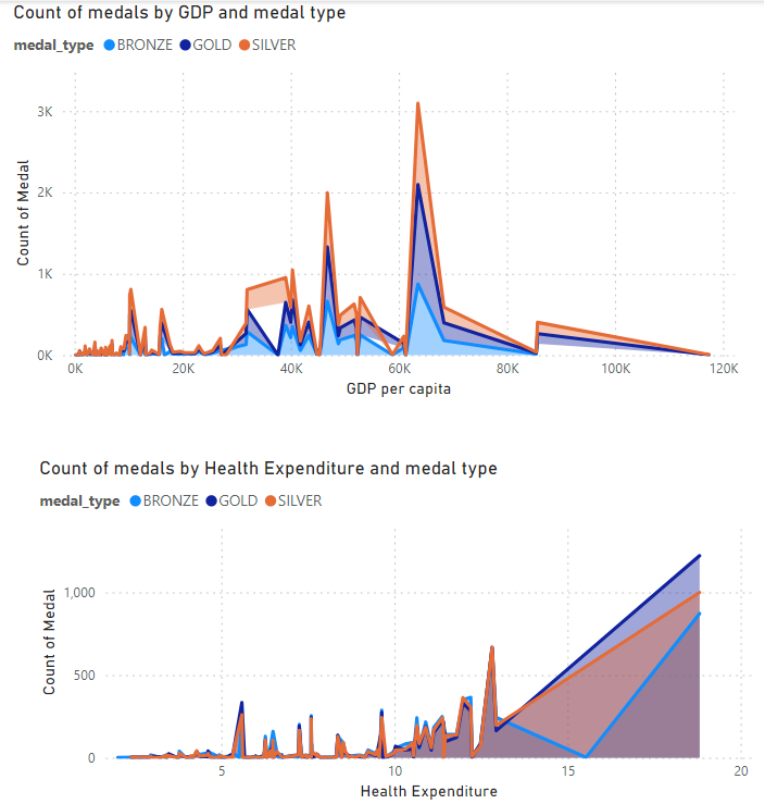
2. **GDP and Olympic Success:** There appears to be a correlation between a country's GDP and its Olympic medal tally, with countries in the GDP range of 60-80k showcasing the highest medal counts. Interestingly, medal counts fluctuate between different GDP brackets and decline after reaching the peak bracket. However, an increase in health expenditure seems to positively influence the number of medals won.
3. **Trends in Athlete Participation:** The Tokyo 2020 Olympics, followed by Rio 2016 and Beijing, have recorded the highest athlete participation. The data also indicates that host nations such as the US, Japan, and Great Britain have secured the most medals.
4. **Continental Medal Trends:** Over recent decades, Europe, North America, and Asia stand out as the continents with the highest medal counts. A clear upward trend in medals won is evident, with notable periodic fluctuations detected since the 1980s.
5. **Seasonal Effects on Medal Distribution:** Seasonal variations significantly impact medal distribution, with a higher frequency of medals won during the summer. This seasonal effect is evident in time-series visualizations, highlighting the difference in performance.

Query 1:

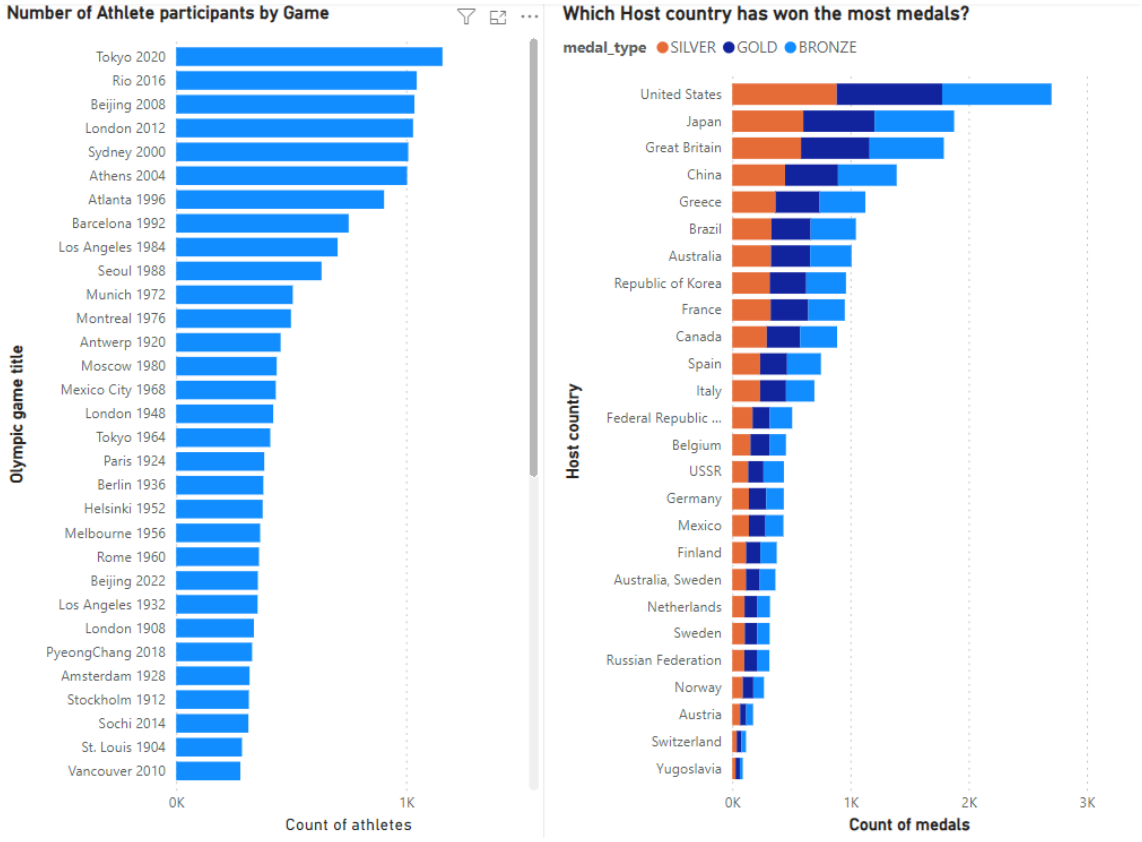


discipline	Sum of gold_medals	Sum of silver_medals	Sum of bronze_medals
Athletics	942	951	932
Swimming	568	557	556
Wrestling	352	375	454
Rowing	333	328	333
Shooting	259	268	262
Sailing	254	249	245
Boxing	242	234	426
Gymnastics Artistic	236	232	261
Canoe Sprint	212	241	255
Fencing	201	206	204
Cycling Track	184	185	182
Speed skating	180	189	181
Weightlifting	172	186	210
Cross Country Skiing	164	165	163
Total	**6406**	**6456**	**6930**

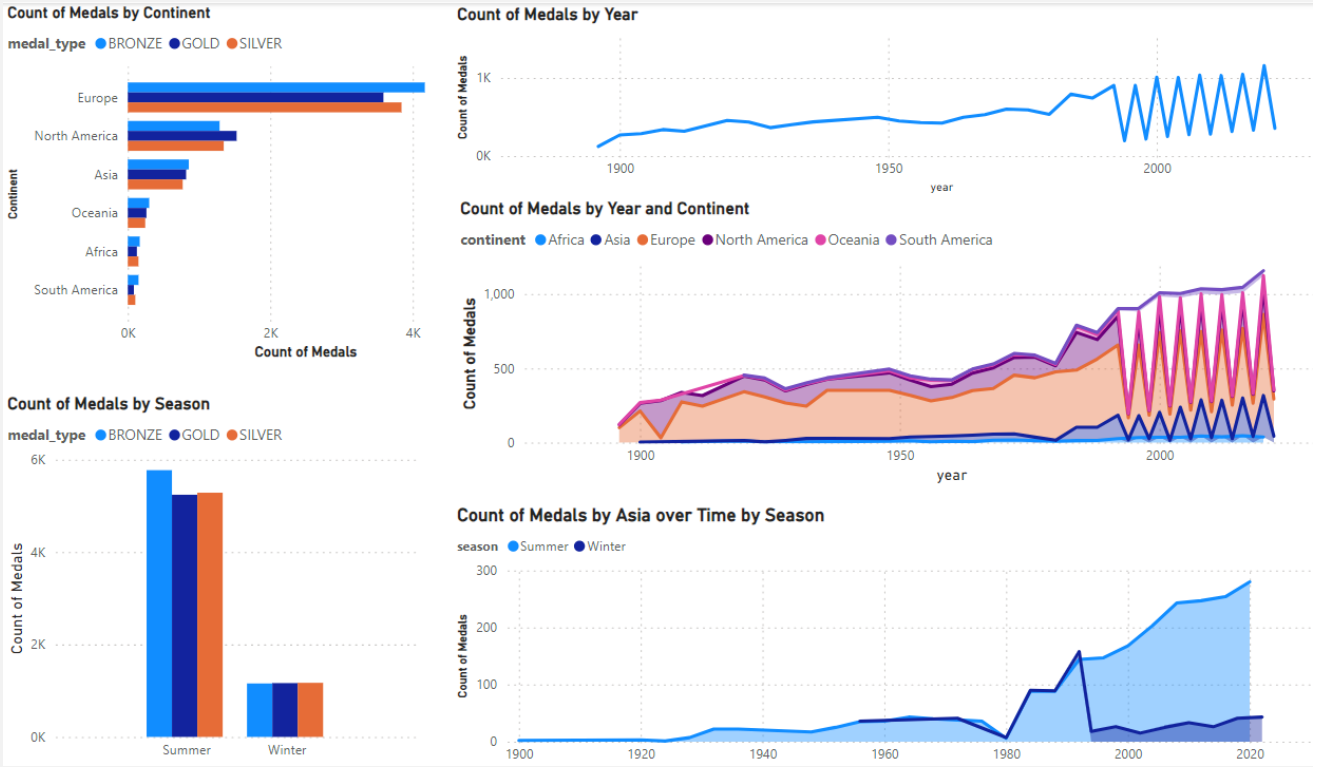
Query 2:



Query 3:



Query 4 and 5:



Argumentative Essay

The Transition from OLAP Database Cubes to Modern Data Technologies

The modern field of data analytics is witnessing a pivotal shift from traditional database cubes to more advanced and efficient technologies that can withhold modern data demands. Albert Wong's 2023 article underscores a growing consensus that database cubes, once staples of data warehousing, are giving way to newer technologies. This essay supports Albert Wong's argument and examines the transition from traditional OLAP cubes to advanced solutions like columnar databases and cloud analysis tools, which offer a timely solution to the growing demands for real-time data processing and visualization [5].

Data cubes, or OLAP (Online Analytical Processing) cubes, have been foundational in facilitating multidimensional data analysis. They allow analysts to swiftly summarize, explore, and analyse data across various dimensions, such as time, geography, and product categories [3]. However, as Wong (2023) and further supported by industry insights [3,4], the limitations of OLAP cubes in scalability, real-time processing, and flexibility are increasingly apparent in the face of growing data complexity and volume.

Columnar databases are one of the few modern technologies that offer substantial improvements over OLAP cubes. These databases store data by columns rather than rows, optimising the speed and efficiency of data retrieval and aggregation which is crucial for handling large amounts of datasets [1]. These databases are adept at handling the enormity of data that modern enterprises generate—a critical need that OLAP cubes cannot handle. Furthermore, columnar structures excel at adapting to the fluid nature of today's data allowing for real-time analysis and decision-making, while traditional database cubes cannot. This refutes the lingering belief that the structured querying ability of OLAP cubes is adequate for contemporary data challenges.

Moreover, cloud-native analytics is another modern tool that provide scalability and flexibility, enabling businesses to adapt to data needs in real-time with increased efficiency and low cost. This tool has the ability to deliver instantaneous data analysis, a feature that OLAP cubes were not designed to perform. The argument that OLAP cubes can function offline, is now trivial in our interconnected reality. Modern businesses operate in a realm of constant connectivity, where real-time access to data analytics is not just preferred but expected. Hence, the offline capabilities of OLAP cubes no longer constitute a significant advantage.

The debate is not just about choosing between database cubes and newer technologies but rather about understanding the specific needs of modern businesses that demand real-time insights and high scalability. Columnar databases not only provide the requisite speed and efficiency but also reduce the costs and complexity of data management, a significant advantage over OLAP cubes [1]. In-memory analytics, another modern approach, offers an even more compelling argument against OLAP cubes. By storing data in RAM instead of on physical disks, in-memory analytics systems deliver unprecedented speed, facilitating real-time data analysis and decision-making capabilities that OLAP cubes cannot match [4].

The transition from OLAP cubes to modern technologies such as columnar databases and in-memory analytics is not just inevitable but essential for businesses aiming to leverage big data effectively. Organizations should embrace these advancements to stay competitive, ensuring agility and efficiency in their data analysis operations.

References:

1. *Data Cube vs. Data Warehouse for Business Intelligence | Sigma Computing*. (n.d.). Www.sigmacomputing.com. https://www.sigmacomputing.com/blog/data-cube-vs-data-warehouse-for-business-intelligence
2. *Is OLAP Dead?* (n.d.). KDnuggets. Retrieved April 14, 2024, from https://www.kdnuggets.com/2022/10/olap-dead.html
3. The Rise and Fall of the OLAP Cube. (2020, January 30). The Holistics Blog. https://www.holistics.io/blog/the-rise-and-fall-of-the-olap-cube/#:~:text=One%20of%20the%20biggest%20shifts
4. OLAP cubes, outdated BI technology? (2010, October 14). Yellowfin BI. https://www.yellowfinbi.com/blog/olap-cubes-outdated-bi-technology
5. Wong, A. (2024, March 22). *Database cubes are dead; what is their replacement?* Medium. https://atwong.medium.com/database-cubes-are-dead-what-is-their-replacement-999a0014f32c

Associate Rule Mining

Objective

The primary objective of this analysis is to use Association Rule Mining to uncover significant patterns within the Olympic Games dataset. This process helps in identifying relationships between different attributes such as athlete nationality, medal type, gender, and event disciplines. Understanding these patterns can provide actionable insights for commercial strategies.

Methodology

The analysis was performed using Python libraries such as pandas, NumPy, and mlxtend. The dataset included multiple dimensions like athlete information, event details, medal records, and economic indicators associated with the games. After preprocessing, the data was transformed into a format suitable for mining by using the TransactionEncoder from mlxtend, which converts the dataset into a one-hot encoded boolean matrix. The Apriori algorithm was then applied to find frequent itemsets, followed by generating association rules with metrics like support, confidence, and lift.

Key Findings

The top association rules identified were primarily around the types of medals won by male athletes across different events. For instance, significant rules included:

1. **(BRONZE) -> (Men)** with a support of 0.226990, confidence of 0.654137, and a lift of 1.007226.
2. **(GOLD) -> (Men)** with a support of 0.212149, confidence of 0.647489, and a lift of 0.996989.
3. **(SILVER) -> (Men)** with a support of 0.210306, confidence of 0.646409, and a lift of 0.995326.

These rules indicate a high probability that if an athlete wins a medal, and if the medal is bronze, gold, or silver, the athlete is likely to be male. This suggests a dominance or higher participation rate of male athletes in medal-earning positions.

Insights and Recommendations

The association rules provide insights into gender disparities and event-specific performance, which can guide marketing and strategic decisions. Based on the findings, the following commercial suggestions are offered:

1. **Gender-focused Initiatives**: Given the dominance of male athletes in winning medals, there could be a need to promote and support female participation and success in the Olympics through targeted training programs and marketing campaigns.
2. **Event-specific Promotions**: Understanding which events consistently see success from certain countries or genders can help in tailoring promotions and sponsorships. For instance, countries leading in specific events might be targeted for event-specific merchandise or advertising campaigns.
3. **Athlete Endorsements**: Athletes who frequently win medals present a lucrative opportunity for brand endorsements. By focusing on athletes with high confidence and support in winning medals, brands can strategically align with winners, enhancing both athlete and brand visibility.

Potential Limitations

The analysis might not have uncovered meaningful rules across all expected dimensions due to limitations in the dataset's diversity or completeness. Additionally, the support threshold might have been too high to capture more nuanced relationships, or the dataset size and imbalance between categories could have influenced the results.