REPORT: DATA WAREHOUSE DESIGN

[Jaya Motwani: 23990486]

Executive Summary

An overview of the data warehouse design solution provided, highlighting key insights and strategic recommendations derived from the data.

Introduction

Explanation of the report's objectives, scope, and the specific business clients and queries addressed.

Client Analysis

Business Clients

- **Identification of Key Clients**: Profiling two primary business clients, explaining their significance and role in the data analysis context.
- **Client-Centric Queries**: Elaboration on the ten queries of interest and their relevance to the business objectives.

Data Modeling and Star Schema Design

Star Net Diagram

- **Conceptual Design**: Presentation of the star net diagram that illustrates the conceptual design based on the lowest level of information access.
- **Dimensions and Hierarchies**: Identification of the dimensions and concept hierarchies that align with the star net design.
- **Schema and Hierarchies Description**: Detailed schema of each dimension, along with the concept hierarchies.

Star Schema (Data Warehouse Design)

- **Fact and Dimension Tables**: Definitions and schematics of fact tables, dimension tables, and their relationships.
- **Schema Visualization**: An entity-relationship (ER) diagram visualizing the star schema.
- **Grain of Fact Table**: Description of the granularity of the fact table and its implications for analysis.
- **Measures and Dimensions**: Specification of the measures and dimensions, including dimension references.

ETL Process

- **Data Preparation**: Outline of the data cleaning, pre-processing, and ETL (Extract, Transform, Load) process.
- **Transformation Details**: Detailed description of the data transformation steps with supporting code, screenshots, or explanatory text.

Multidimensional Analysis

- **Cubes and OLAP Operations**: Explanation of how multidimensional cubes support analytical operations such as roll-up and drill-down.

Data Visualization with Atoti

- **Atoti Visualization**: Demonstration of query result visualization using the Atoti platform.

Association Rule Mining

- **Process Description**: Detailed description of the association rule mining process and the methodology employed.
- **Top K Rules**: Presentation of the top K rules discovered (with $K \geq 1$) and their significance.
- **Rule Insights**: Interpretation of insights gained from the rule mining results.
- **Commerce Suggestions**: Provision of at least three business suggestions based on the results of the association rule mining.

This outline provides a structured approach to report the findings of my data warehouse project from the identification of business clients and queries through to data modeling, the ETL process, OLAP operations, visualization, and finally, the application of association rule mining. It serves as a detailed roadmap for the report's readers, setting clear expectations for the insights and recommendations contained within.

References

Dataset and Problem Domain

The Olympic Games represent the sole global, multi-disciplinary sports event, celebrated worldwide. Featuring participation from over 200 nations in more than 400 events spanning both the Summer and Winter Games, the Olympics serve as a platform for global competition, inspiration, and unity.

Modelling Business Process

Client Identification and Business Queries

The two potential clients identified for modelling business process related to Olympic dataset with specific focus are as follows:

I. **Athletics Performance Analyst Company (Client A)**: This client is primarily interested in analyzing and improving the performance of countries in athletics events at the Olympics. They are interested in understanding the performance trends based on factors such as gender, regions, historical performance, and other relevant demographic and socio-economic indicators. Their goal is to provide insights and recommendations to countries, athletic organizations, and coaches to optimize their training programs, talent identification strategies, and resource allocation. This client might be interested in seeking insights on following issues:

   1) Identify the top 10 performing countries in Athletics and provide a breakdown of these medals by gender.
   2) Analyze and compare the performance of different regions in various gender categories across all events to determine which region has performed the best in Athletics.
   3) Discover insights into the distribution and trends of marathon medals, across all regions and years to predict future potential medal winners based on historical data.
   4) Analyze medal distributions across different events in women's athletics to identify patterns, strengths, and areas for improvement in athletic performance and training.
   5) Analyze performance of the United States in Olympics during the year 2018, focusing on the number of medals won and the disciplines in which these medals were achieved.

II. **Health and Well-being Research Institute (Client B)**: This client is interested in exploring the relationship between Olympic performance and various health-related indicators, such as life expectancy and depression rates. They aim to uncover potential correlations and causal relationships between participating in the Olympics, achieving success, and the overall health and well-being of populations in different countries. Their objective is to conduct research that informs policies and interventions aimed at promoting physical and mental health, particularly among athletes and the broader population.

   1) Provide a trend analysis on participant demographics across various sporting events in Olympics from 1992 to 2018, which can then be leveraged for business analysis and strategic planning in the sports industry.
   2) Could you analyze and compare the impact levels of depression among athletes from the top 10 performing countries in the 2018 Olympics?
   3) Could you analyze and compare the impact levels of depression among athletes from the bottom 10 performing countries in the 2018 Olympics?
   4) Could you provide an analysis on how life expectancy correlates with Olympic performance across different regions?
   5) Could you provide an analysis on how impact of depression levels correlates with Olympic performance across different regions?

These two clients have distinct focuses and objectives based on their areas of expertise and interest. By leveraging an Olympic dataset, they can gain valuable insights relevant to their respective domains and make informed decisions to drive improvements in athletics performance and population health outcomes.

Star Net Footprints

To answer above queries, following star net footprints have been designed:

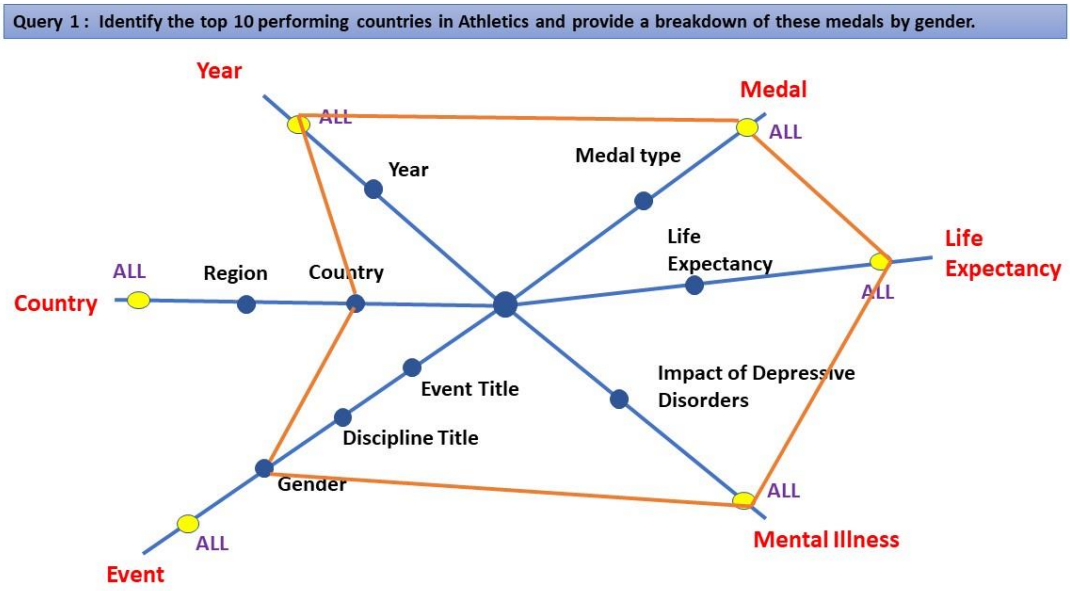


Figure 1: Star Net Footprint (Query 1)

Explanation: Figure 1 presents a schema comprising six distinct dimensions: Year, Medal, Life Expectancy, Country, Event, and Mental Illness. Each dimension is structured with hierarchical levels that facilitate the extraction of detailed insights. *To address Query 1, which requires an analysis of medal distribution by gender within the discipline of Athletics across different countries, we engage with specific segments of our schema. We utilize the Country dimension in conjunction with the Gender level (nested within the Gender -> Discipline hierarchy) while keeping the other dimensions at their highest aggregation level*. This approach allows us to synthesize the necessary information effectively. Thus, the schema's star net architecture is instrumental in navigating and fulfilling the requirements of complex business inquiries.

Few more examples of star net footprints for addressing other queries are given below:

Query 3 : Discover insights into the distribution and trends of marathon medals, across all regions and years to predict future potential medal winners based on historical data.

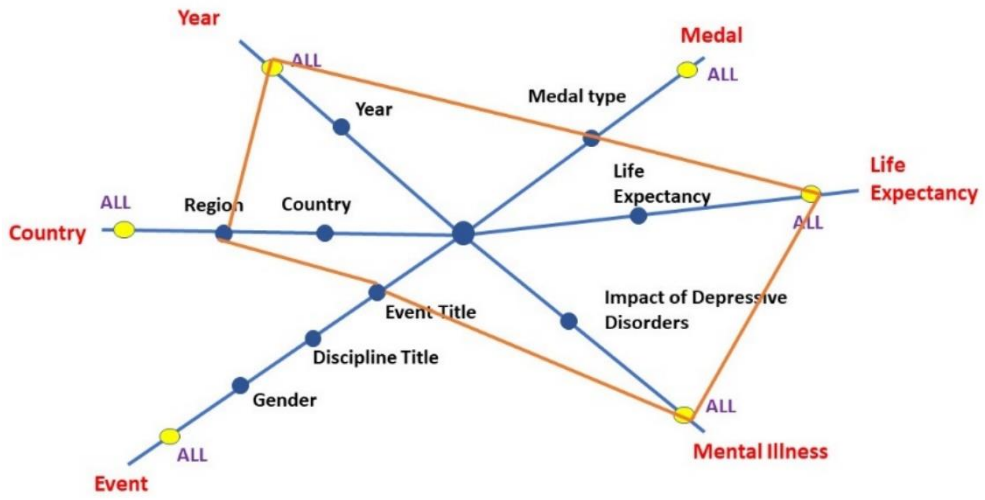


Figure 2: Star Net Footprint (Query 2)

Query 5 : Analyze performance of the United States in Olympics during the year 2018, focusing on the number of medals won and the disciplines in which these medals were achieved.

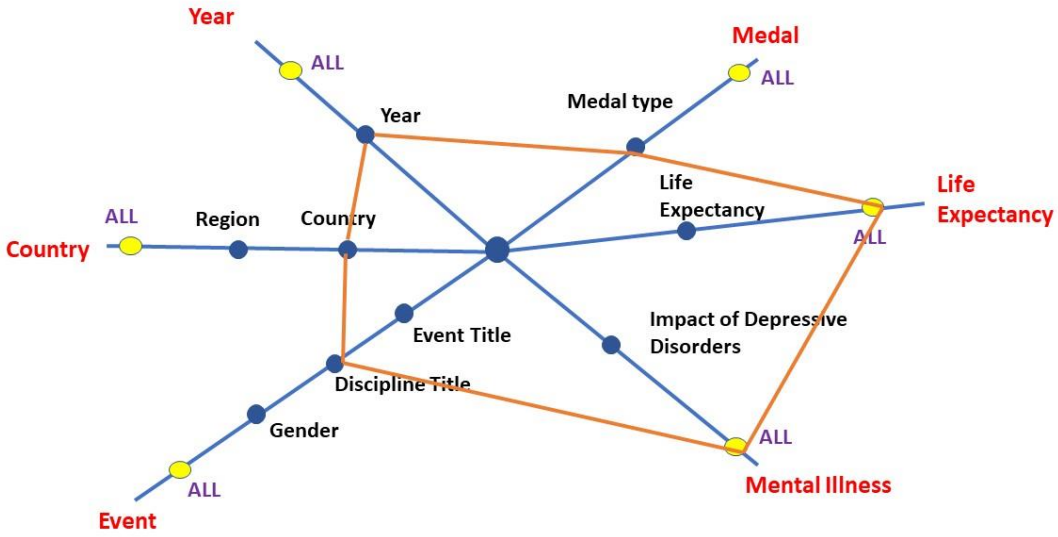


Figure 3: Star Net Footprint (Query 5)

Query 7 : Could you analyze and compare the impact levels of depression among athletes from the top 10 performing countries in the 2018 Olympics?

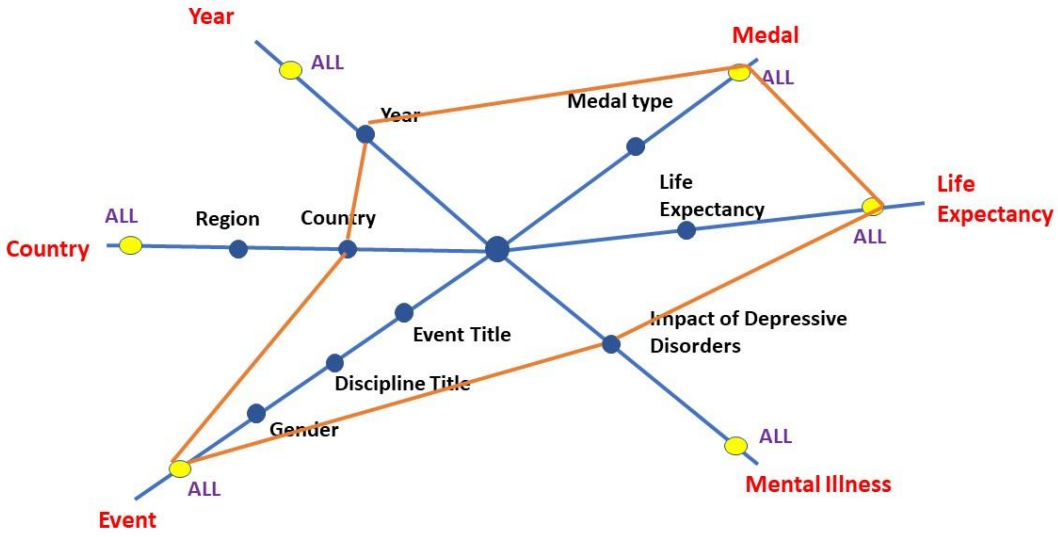


Figure 4: Star Net Footprint (Query 7)

Dimensions and Concept Hierarchies:

The data warehouse incorporates six core dimensions: Country, Event, Life Expectancy, Medal, Mental Illness, and Year. These dimensions, with their respective hierarchies, allow for a comprehensive analysis ranging from geographical and temporal trends to health and event-specific details.

1. **Country: Region --> Country**
2. **Event: Event Gender → Discipline Title → Event Title**
3. **Life Expectancy: Life Expectancy**
4. **Medal: Medal Type**
5. **Mental Illness: DALYs impact levels for depression**
6. **Year: Year**

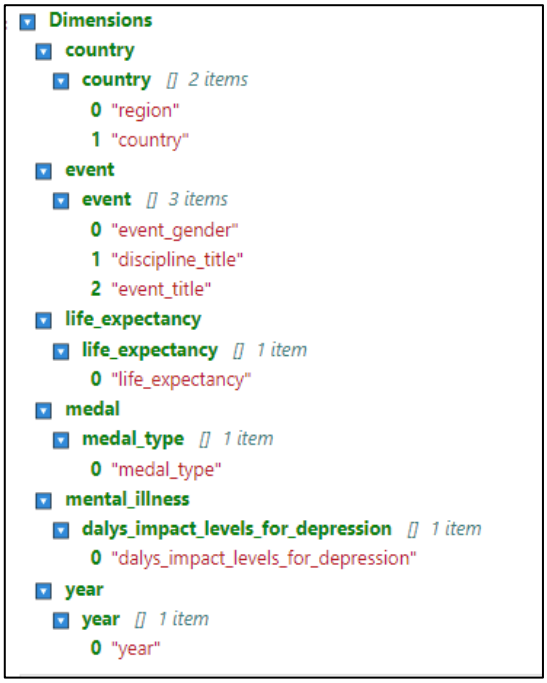


Figure 5: Dimensions and Concept Hierarchies

Data warehouse Design (STAR SCHEMA)

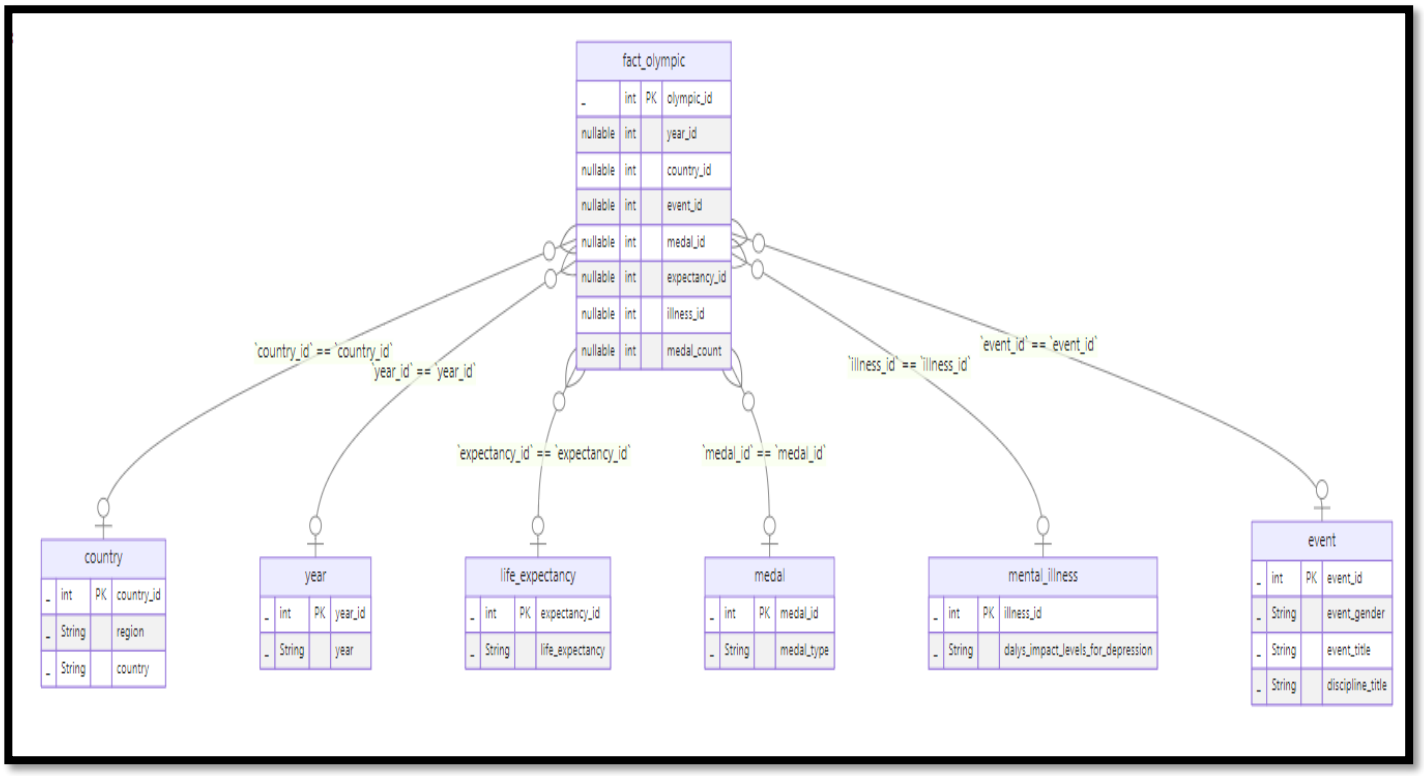


Figure 6: STAR Schema / ER Diagram

Explanation of Star Schema: The central fact table would likely contain records of events, performances, or observations that pertain to the subjects of the dimensions (such as athletic events or health surveys). Each record in this fact table would have foreign keys that reference entries in the six dimension tables. Here's a brief description of how these tables are structured:

Factless Fact Table:

- **Contains**: Dummy measure that is typically numerical and additive, such as the number of medals count (equal to 1 for all rows)
- **Linked by**: Foreign keys that correspond to the primary keys in the dimension tables.

Dimension Tables:

- **Country**: Includes attributes such as country name and region. It allows users to filter and group data by geographical areas.
- **Event**: Contains information about the event, which might include gender categories, discipline titles, and specific event titles.
- **Life Expectancy**: Stores data related to life expectancy, potentially broken down into various categories or groups.
- **Medal**: Holds details about the medals awarded, such as medal type (Gold, Silver, Bronze).
- **Mental Illness**: Could contain measures of the impact of mental illnesses like depression, with DALYs (disability-adjusted life years) as a possible metric.
- **Year**: Lists years and is typically used to analyze trends over time.

Dimension Tables (Detailed Schema)

	country_id	region	country
0	1	Asia	India
1	2	Asia	China
2	3	North America	United States
3	4	Asia	Indonesia
4	5	Asia	Pakistan

Dim table 1: Country
(Reference: country_id)

	year_id	year
0	1	2022
1	2	2020
2	3	2018
3	4	2016
4	5	2014

Dim table 2: Year
(Reference: year_id)

	expectancy_id	life_expectancy
0	1	Low
45	2	Medium-Low
166	3	Medium-High
182	4	High

Dim table 3: Life_Expectancy
(Reference: expectancy_id)

	illness_id	dalys_impact_levels_for_depression
0	1	Critical
60	2	Minimal
90	3	Severe
157	4	Moderate

Dim table 4: Mental_Illness
(Reference: illness_id)

	event_id	event_gender	event_title	discipline_title
0	1	Mixed	Mixed Doubles	Curling
6	2	Women	Women	Curling
9	3	Men	Men	Curling
12	4	Men	Men's Moguls	Freestyle Skiing
15	5	Men	Men's Freeski Halfpipe	Freestyle Skiing

Dim table 5: Event
(Reference: event_id)

	medal_id	medal_type
0	1	GOLD
2	2	SILVER
4	3	BRONZE

Dim table 6: Medal
(Reference: medal_id)

Factless Fact Table

	olympic_id	year_id	country_id	event_id	medal_id	expectancy_id	illness_id	medal_count
0	1	8	25	59	3	4	4	1
1	2	8	32	88	2	4	3	1
2	3	8	35	176	3	4	2	1
3	4	14	25	59	1	4	3	1
4	5	16	88	233	2	4	3	1

Fact table: Olympic_Medal
Pk: Olympic_id
Dummy Measure: medal_count

Grain of the fact table = "An Olympic event with country participating along with health indicator details."

	year	region	country	discipline_title	event_title	event_gender	medal_type	life_expectancy	dalys_impact_levels_for_depression
0	2008	Europe	Italy	Fencing	épée team men	Men	BRONZE	High	Moderate
1	2008	Europe	Spain	Synchronized Swimming	team women	Women	SILVER	High	Severe
2	2008	South America	Argentina	Basketball	basketball men	Men	BRONZE	High	Minimal
3	1996	Europe	Italy	Fencing	épée team men	Men	GOLD	High	Severe
4	1992	Europe	Sweden	Handball	handball men	Men	SILVER	High	Severe

Figure 7: Grain of a fact table

Explanation: The grain of the fact table means that each record represents a unique combination of these attributes as shown in Figure 7, meaning that for every Olympic event, the detailed outcomes (medal type), along with the health indicators (life expectancy and mental health impact), are associated with a specific country and year. This granularity allows for a precise analysis of how participation in Olympic events correlates with health factors across different countries and times.

ETL Process

The ETL process followed for this modelling is shown in Figure 8:

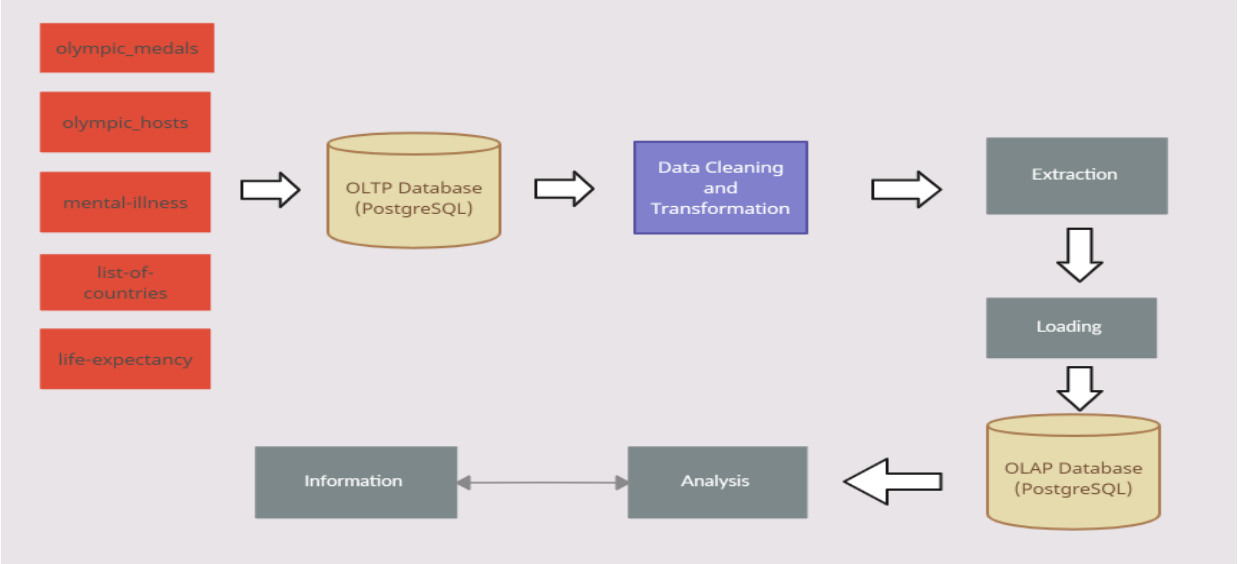


Figure 8: ETL Process

Detailed Explanation of ETL (Extract, Transform, Load) process:

(Code snippets attached)

1. **Extract (E)**:

- Extracted data from provided CSV files:

```
life_expectancy_raw = pd.read_csv('Olympic/life-expectancy.csv')
country_raw = pd.read_csv('Olympic/list-of-countries_areas-by-continent-2024.csv')
mental_illness_raw = pd.read_csv('Olympic/mental-illness.csv')
olympic_hosts_raw = pd.read_csv('Olympic/olympic_hosts.csv')
olympic_medals_raw = pd.read_csv('Olympic/olympic_medals.csv')
```

- Created OLTP Database and table schemas in PostgreSQL.

```
# Only creating the database if it does not exist
if not db_exists:
    sql_create_db = "CREATE DATABASE oltp_db"
    cursor.execute(sql_create_db)
    print("Database created successfully........")
else:
    print("Database already exists.")
```

- Extracted the raw data into tables created in OLTP Database in PostgreSQL.

```
Data from life_expectancy_oltp:
('Afghanistan', 'AFG', 1950, 27.7275)
('Afghanistan', 'AFG', 1951, 27.9634)
('Afghanistan', 'AFG', 1952, 28.4456)
Data from country_oltp:
('India', 'Asia')
('China', 'Asia')
('United States', 'North America')
Data from mental_illness_oltp:
('Afghanistan', 'AFG', 1990, 895.22565, 138.24825, 147.64412, 26.471115, 440.33)
('Afghanistan', 'AFG', 1991, 893.88434, 137.76122, 147.56696, 25.548681, 439.47202)
('Afghanistan', 'AFG', 1992, 892.34973, 137.0803, 147.13086, 24.637949, 437.60718)
Data from olympic_hosts_oltp:
('beijing-2022', datetime.date(2022, 2, 20), datetime.date(2022, 2, 4), 'China', 'Beijing 2022', 'Winter', 2022)
('tokyo-2020', datetime.date(2021, 8, 8), datetime.date(2021, 7, 23), 'Japan', 'Tokyo 2020', 'Summer', 2020)
('pyeongchang-2018', datetime.date(2018, 2, 25), datetime.date(2018, 2, 8), 'Republic of Korea', 'PyeongChang 2018', 'Winter', 2018)
Data from olympic_medals_oltp:
('Curling', 'beijing-2022', 'Mixed Doubles', 'Mixed', 'GOLD', 'GameTeam', 'Italy', 'https://olympics.com/en/athletes/stefania-constantini', 'Stefania CONSTANTINI', 'Italy', 'IT', 'ITA')
('Curling', 'beijing-2022', 'Mixed Doubles', 'Mixed', 'GOLD', 'GameTeam', 'Italy', 'https://olympics.com/en/athletes/amos-mosaner', 'Amos MOSANER', 'Italy', 'IT', 'ITA')
('Curling', 'beijing-2022', 'Mixed Doubles', 'Mixed', 'SILVER', 'GameTeam', 'Norway', 'https://olympics.com/en/athletes/kristin-skaslien', 'Kristin SKASLIEN', 'Norway', 'NO', 'NOR')
```

- Loaded the data from PostgreSQL tables in OLTP Database to data frames using Pandas for pre-processing:

```
# Setting up database connection using SQLAlchemy
engine = create_engine('postgresql://postgres:postgres@localhost:5432/oltp_db')

# Table names to DataFrame dictionary
tables = {
    'country_oltp': 'country_df',
    'life_expectancy_oltp': 'life_expectancy_df',
    'mental_illness_oltp': 'mental_illness_df',
    'olympic_hosts_oltp': 'olympic_hosts_df',
    'olympic_medals_oltp': 'olympic_medals_df'
}

# Loading each table into a DataFrame
for table_name, df_name in tables.items():
    query = f"SELECT * FROM {table_name};"
    locals()[df_name] = pd.read_sql_query(query, engine)
    print(f"Data from {table_name} loaded into DataFrame '{df_name}'.")


# Closing the database connection
engine.dispose()
```

```
Data from country_oltp loaded into DataFrame 'country_df'.
Data from life_expectancy_oltp loaded into DataFrame 'life_expectancy_df'.
Data from mental_illness_oltp loaded into DataFrame 'mental_illness_df'.
Data from olympic_hosts_oltp loaded into DataFrame 'olympic_hosts_df'.
Data from olympic_medals_oltp loaded into DataFrame 'olympic_medals_df'.
```

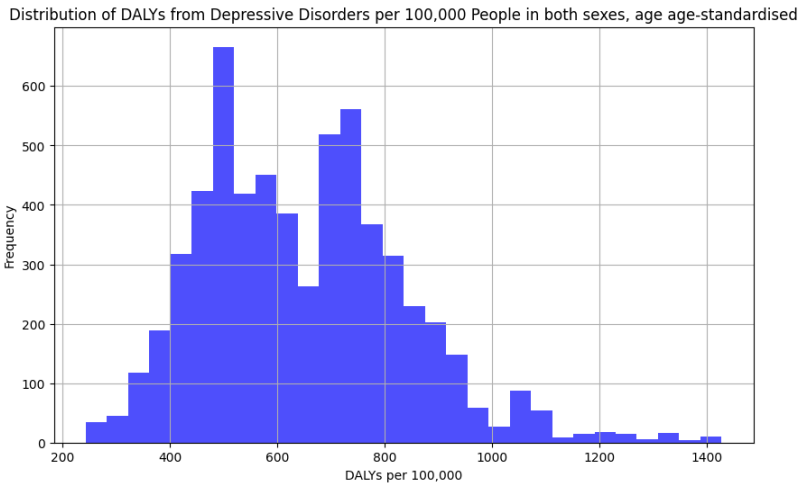
2. **Transform (T)**:

- Matched Country names across all tables and applied following country mapping:

```
country_name_mapping = {
    "Soviet Union": "Russia",
    "Unified Team": "Russia",
    "West Germany": "Germany",
    "East Germany": "Germany",
    "Great Britain": "United Kingdom",
    "United States of America": "United States",
    "People's Republic of China": "China",
    "ROC": "Russia",
    "Czech Republic": "Czechia",
    "Hong Kong, China": "Hong Kong",
}
```

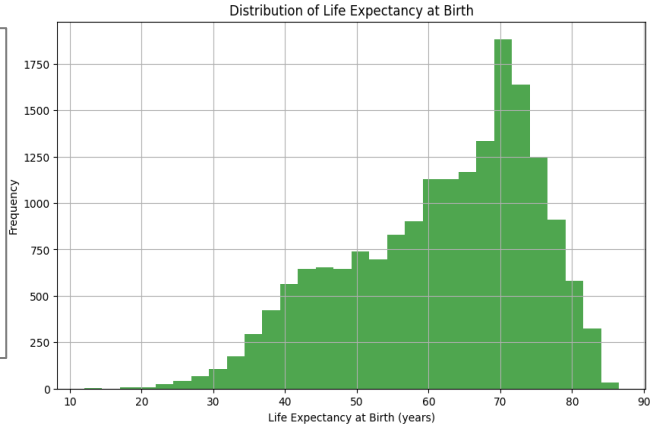
- Categorized DALYs from Depressive disorders based on histogram:

```
def categorize_dalys(value):
    if value <= quartiles[0.25]:
        return 'Minimal'
    elif value <= quartiles[0.50]:
        return 'Moderate'
    elif value <= quartiles[0.75]:
        return 'Severe'
    else:
        return 'Critical'
```



- Categorized life expectancy at birth based on histogram:

```
# Function to categorize life expectancy based on quartiles
def categorize_life_expectancy(value, quartiles):
    if value <= quartiles[0.25]:
        return 'Low'
    elif value <= quartiles[0.50]:
        return 'Medium-Low'
    elif value <= quartiles[0.75]:
        return 'Medium-High'
    else:
        return 'High'
```



- Dropped duplicates, renamed columns and added primary key to all columns.

3. **Load (L):**

- Created of OLAP database and all dimension and fact table schemas in PostgreSQL.

```
# Only create the database if it does not exist
if not db_exists:
    sql_create_db = "CREATE DATABASE olap_db"
    cursor.execute(sql_create_db)
    print("Database created successfully........")
else:

    print("Database already exists.")
```

- Loaded data into all tables in OLAP database.

```
# Create SQLAlchemy engine
engine = create_engine('postgresql://postgres:postgres@localhost:5432/olap_db')

# Insert data into PostgreSQL
year.to_sql('dim_year', engine, if_exists='append', index=False)
print("Data inserted into year")

country.to_sql('dim_country', engine, if_exists='append', index=False)
print("Data inserted into country")

event.to_sql('dim_event', engine, if_exists='append', index=False)
print("Data inserted into event")

medal.to_sql('dim_medal', engine, if_exists='append', index=False)
print("Data inserted into medal")

mental_illness.to_sql('dim_mental_illness', engine, if_exists='append', index=False)
print("Data inserted into mental_illness")

life_expectancy.to_sql('dim_life_expectancy', engine, if_exists='append', index=False)
print("Data inserted into life_expectancy")

merged_df.to_sql('fact_olympic_medal', engine, if_exists='append', index=False)
print("Data inserted into fact_olympic_medal")
```

```
Data inserted into year
Data inserted into country
Data inserted into event
Data inserted into medal
Data inserted into mental_illness
Data inserted into life_expectancy
Data inserted into fact_olympic_medal
```

Querying and Visualisation

(Please refer to visuals of query no 2A, 2B and 4 for cube roll up and drill down operations)

Cube Roll-Up: This has been demonstrated in visuals of **query no 2A and 2B**. The roll-up operation summarizes or aggregates data, increasing the level of abstraction. This can involve ascending up a hierarchy (e.g., from event gender to all genders, from specific region to all regions) or reducing dimensions from the cube.

Cube Drill-Down: This has been demonstrated in visuals of **query no 4**. The drill-down operation provides a more detailed view of the data by descending down the hierarchy or adding more dimensions to the analysis. It allows users to navigate from summary data to more detailed data. In query no 4, a drill-down is illustrated by navigating from one region to all countries in that region.

Query 1: Identify the top 10 performing countries in Athletics and provide a breakdown of these medals by gender.

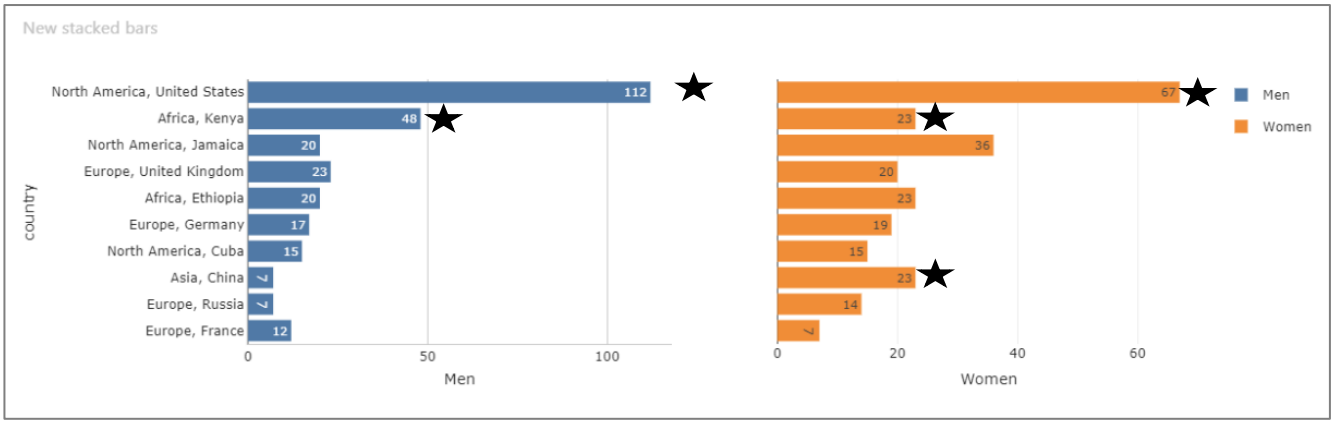
United States tops the list by large margin with all other countries.

Kenya stands at 2nd position just after United States.

In Asia, women have secured more medals than men.

event_gender	region	country	medal_count.SUM
Men	North America	United States	112
	Africa	Kenya	48
	Europe	United Kingdom	23
	Africa	Ethiopia	20
	North America	Jamaica	20
	Europe	Germany	17
	North America	Cuba	15
	Europe	France	12
		Poland	12
	Africa	Morocco	11

event_gender	region	country	medal_count.SUM
Women	North America	United States	67
		Jamaica	36
	Africa	Ethiopia	23
		Kenya	23
	Asia	China	23
	Europe	United Kingdom	20
		Germany	19
	North America	Cuba	15
	Europe	Russia	14
		Ukraine	13



Query 2: Analyze and compare the performance of different regions in various gender categories across all events to determine which region has performed the best in Athletics.

```
query_2 = cube.query(
    measures["medal_count.SUM"],
    levels=[(levels['event', 'event', 'event_gender']),(levels['country','country','region'])],
            filter = (levels['event','event','discipline_title'] == 'Athletics')
           )

# Convert the Atoti query result to a DataFrame if it's not already one
df = pd.DataFrame(query_2)

# Pivot the DataFrame to get regions as columns and event_gender as rows
pivot_table = df.pivot_table(
    values='medal_count.SUM',  # This should match the measure label from your query
    index='event_gender',
    columns='region',
    aggfunc='sum',  # Summing up the medals, adjust if different aggregation is needed
    fill_value=0  # Fills in NaN values with 0 for better presentation
)

pivot_table
```

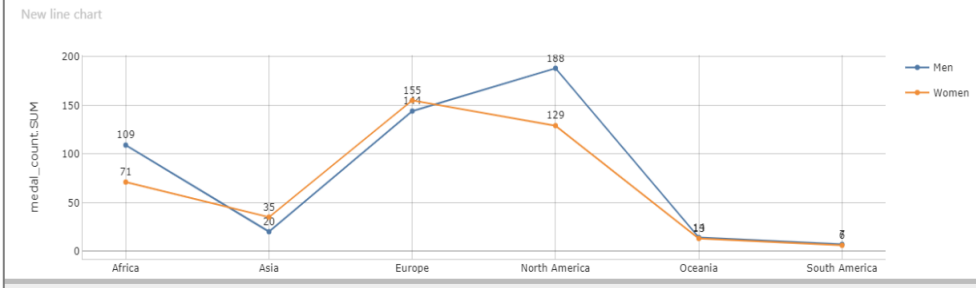
event_gender	Africa	Asia	Europe	North America	Oceania	South America
Men	109	20	144	188	14	7
Women	71	35	155	129	13	6

Interesting trends:

Women performed slightly better than men in Asia and Europe.

In Africa and North America, men outnumbered women by significant number of medal counts.

```
session.widget
```



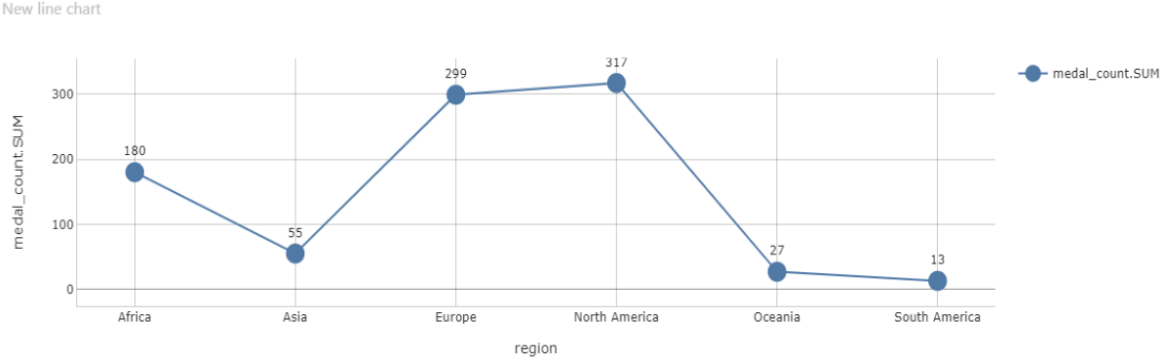
 # Query 2A: Rolling up the cube on Event_Gender dimension to get total medal counts for all regions for all genders in Athletics.

```
query_2A = cube.query(
    measures["medal_count.SUM"],
    levels=[(levels['country','country','region'])],
             filter = (levels['event','event','discipline_title'] == 'Athletics')
            )

# Convert the Atoti query result to a DataFrame if it's not already one
df = pd.DataFrame(query_2A)

# Pivot the DataFrame to get regions as columns and event_gender as rows
pivot_table = df.pivot_table(
    values='medal_count.SUM',  # This should match the measure label from your query
    columns='region',
    aggfunc='sum',  # Summing up the medals, adjust if different aggregation is needed
    fill_value=0  # Fills in NaN values with 0 for better presentation
)

pivot_table
```

region	Africa	Asia	Europe	North America	Oceania	South America
medal_count.SUM	180	55	299	317	27	13

Cube Roll up for total Medal Counts for all genders

```
session.widget
```



 # Query 2B: Rolling up the cube on Country dimension to get total medal counts for all genders for all regions in Athletics

1]:
```
query_2B = cube.query(
    measures["medal_count.SUM"],
    levels=[(levels['event','event','event_gender'])],
             filter = (levels['event','event','discipline_title'] == 'Athletics')
            )

# Convert the Atoti query result to a DataFrame if it's not already one
df = pd.DataFrame(query_2B)

# Pivot the DataFrame to get regions as columns and event_gender as rows
pivot_table = df.pivot_table(
    values='medal_count.SUM',  # This should match the measure label from your query
    columns='event_gender',
    aggfunc='sum',  # Summing up the medals, adjust if different aggregation is needed
    fill_value=0  # Fills in NaN values with 0 for better presentation
)

pivot_table
```

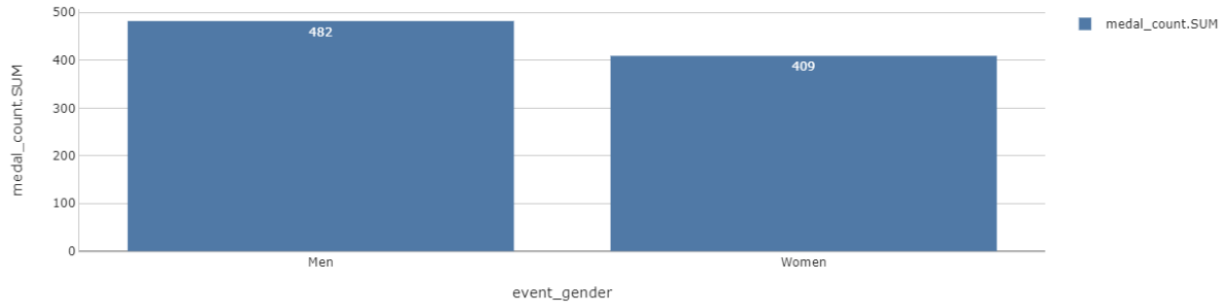
1]:

event_gender	Men	Women
medal_count.SUM	482	409

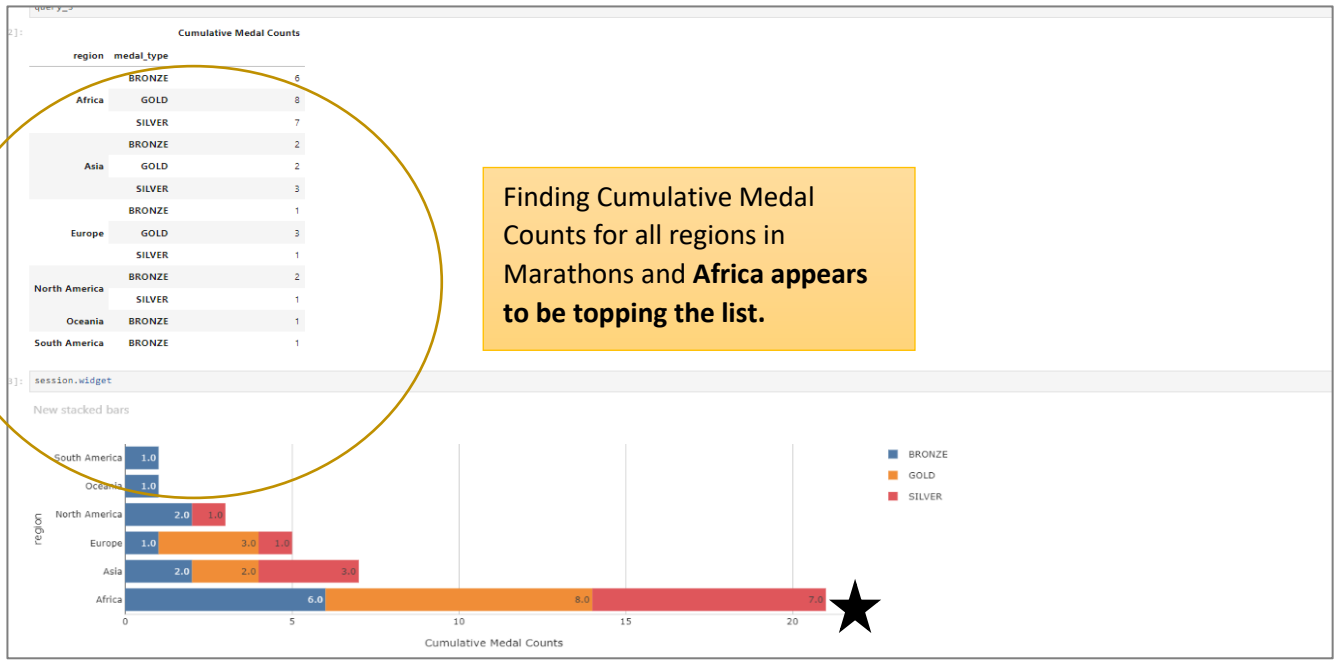
Cube Roll up for total Medal Counts for all regions.

2]:
```
session.widget
```

New stacked columns



Query 3: Discover insights into the distribution and trends of marathon medals, across all regions and years to predict future potential medal winners based on historical data



Cumulative Medal Counts

region	medal_type	
	BRONZE	6
Africa	GOLD	8
	SILVER	7
	BRONZE	2
Asia	GOLD	2
	SILVER	3
	BRONZE	1
Europe	GOLD	3
	SILVER	1
	BRONZE	2
North America	SILVER	1
Oceania	BRONZE	1
South America	BRONZE	1

```
session.widget
```

New stacked bars

Finding Cumulative Medal Counts for all regions in Marathons and **Africa appears to be topping the list.**

 # Query 3A: Drilling down the cube to find out the countries in Africa that have highest Cumulative Medal Counts in Marathon ?

```
query_3A = cube.query(
    measures["Cumulative Medal Counts"],
    levels=[
        levels['country', 'country', 'country'],
        levels['medal', 'medal_type', 'medal_type']
    ],
    filter=(
        (levels['event', 'event', 'event_title'].isin(('marathon men'), ('marathon women'))) &
        (levels['country', 'country', 'region'] == 'Africa')
    )
)
query_3A
```

Cumulative Medal Counts

region	country	medal_type	
		BRONZE	3
	Ethiopia	GOLD	3
		SILVER	1
		BRONZE	3
Africa	Kenya	GOLD	3
		SILVER	5
	Morocco	SILVER	1
	South Africa	GOLD	1
	Uganda	GOLD	1

```
session.widget
```

New stacked bars

Drilling Down the cube to find African countries performing well in Marathons, and **Kenya tops the list** with highest cumulative medal counts, can be a future potential winner.

Query 4: Analyze medal distributions across different events in women's athletics to identify patterns, strengths, and areas for improvement in athletic performance and training

56]:

				medal_count.SUM
event_gender	discipline_title	event_title	medal_type	
Women	Athletics	10000m walk women	BRONZE	2
			GOLD	1
			SILVER	2
		10000m women	BRONZE	7
			GOLD	7
		...	...	...
		shot put women	GOLD	7
			SILVER	6
			BRONZE	5
		triple jump women	GOLD	6
			SILVER	4

75 rows × 1 columns

Atoti visualization in the form of pivot table, analysing medal counts for each medal type and across all events in Athletics discipline for women.

New pivot table

event_gender	discipline_title	event_title	Total medal_count.SUM	BRONZE medal_count.SUM	GOLD medal_count.SUM	SILVER medal_count.SUM
Total			409	139	134	136
Women	Total		409	139	134	136
	Athletics	Total	409	139	134	136
		10000m walk women	5	2	1	2
		10000m women	21	7	7	7
		100m hurdles women	21	7	7	7
		100m women	20	6	6	8
		1500m women	17	6	6	5
		200m women	21	7	7	7

Query 5: Analyze performance of the United States in Olympics during the year 2018, focusing on the number of medals won and the disciplines in which these medals were achieved. The client is interested in United States because as per result of Query 1, United States tops the list of best performers in Athletics in both the gender categories.

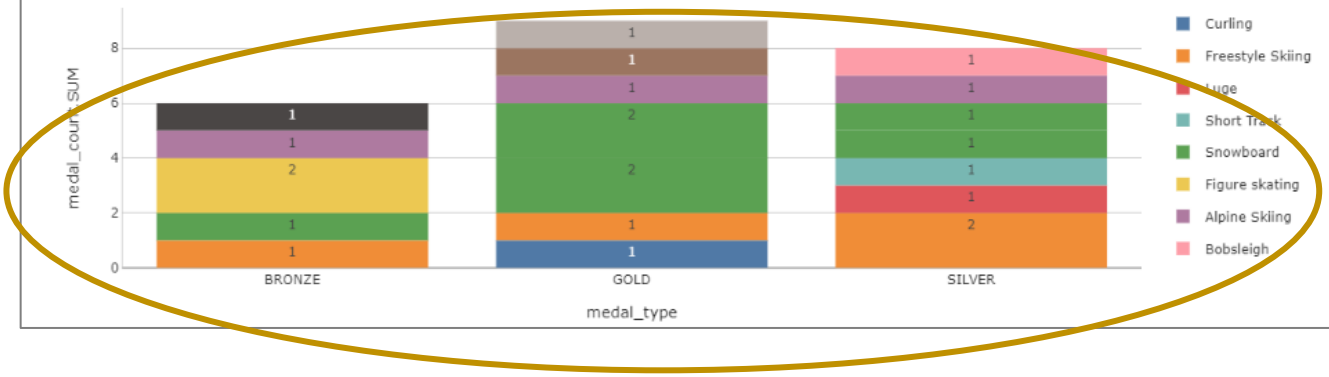
region	country	event_gender	discipline_title	medal_count.SUM	
			Curling	1	
			Freestyle Skiing	3	★
		Men	Luge	1	
			Short Track	1	
			Snowboard	3	★
		Mixed	Figure skating	2	
North America	United States		Alpine Skiing	3	★
			Bobsleigh	1	
			Cross Country Skiing	1	
		Women	Freestyle Skiing	1	
			Ice Hockey	1	
			Snowboard	4	★
			Speed skating	1	

Total Medal Counts for all 3 types of medals won by United States in all events of Athletics in 2018.

Starred disciplines have the highest medal counts for United States.

```
session.widget
```

New stacked columns

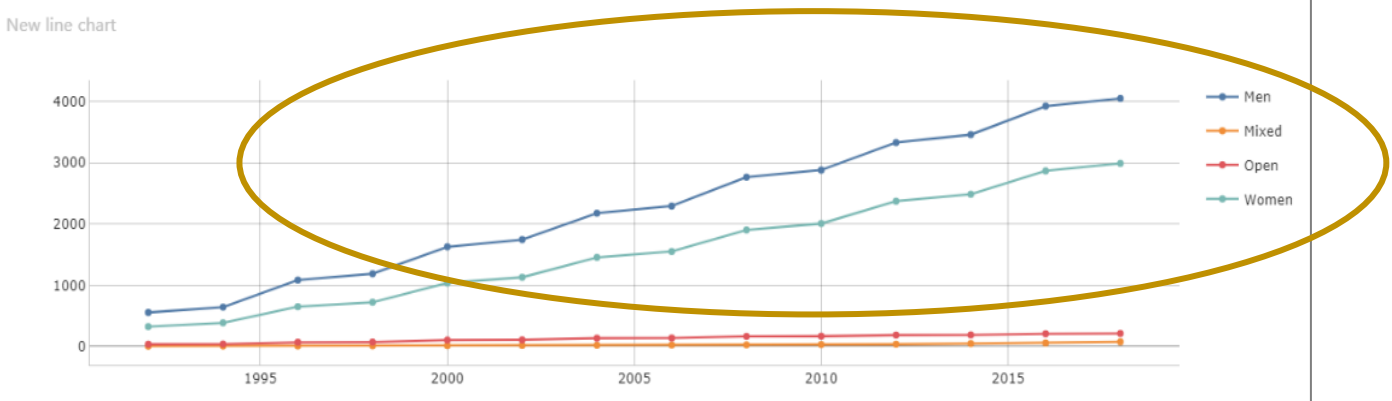


Query 6: Provide a trend analysis on participant demographics across various sporting events in Olympics from 1992 to 2018, which can then be leveraged for business analysis and strategic planning in the sports industry.

event_gender	Men	Mixed	Open	Women
year				
1992	556	6	38	325
1994	643	8	41	387
1996	1084	9	70	650
1998	1187	11	73	723
2000	1627	14	106	1037
2002	1744	18	109	1129
2004	2175	21	136	1454
2006	2291	25	139	1550
2008	2762	27	166	1901
2010	2878	32	169	2006
2012	3327	38	187	2370
2014	3456	48	190	2484
2016	3921	57	208	2865
2018	4046	77	213	2987

Increasing trend year by year highlighting participant demographics i.e., gender profile, for all Olympic events.

```
session.widget
```



Query 7: Could you analyze and compare the impact levels of depression among athletes from the top 10 performing countries in the 2018 Olympics?

region	country	dalys_impact_levels_for_depression	medal_count.SUM
Europe	Norway	Moderate	39
	Germany	Moderate	31
North America	Canada	Moderate	29
	United States	Critical	23
Europe	Netherlands	Moderate	20
	Switzerland	Severe	15
	France	Severe	15
	Austria	Moderate	14
	Sweden	Severe	14
Asia	Japan	Minimal	13

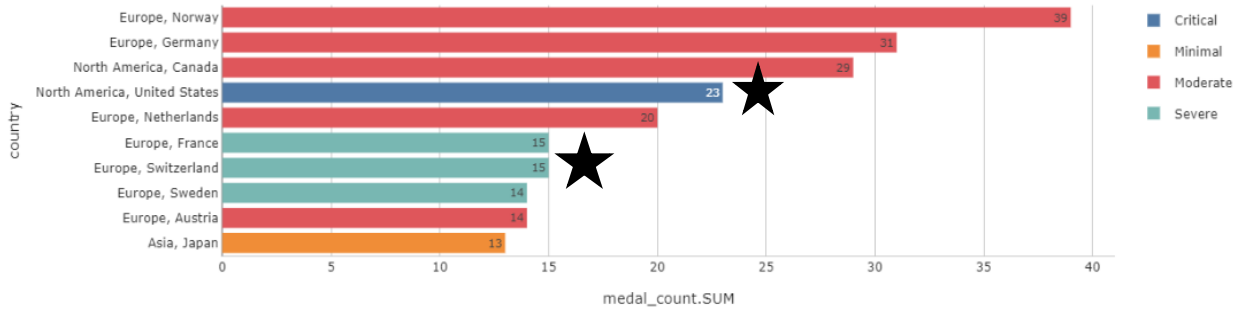
Interesting Trends:

United States – Critical

France, Switzerland, Sweden - Severe

session.widget

New stacked bars



Query 8: Could you analyze and compare the impact levels of depression among athletes from the bottom 10 performing countries in the 2018 Olympics?

region	country	dalys_impact_levels_for_depression	medal_count.SUM
Asia	Kazakhstan	Moderate	1
Europe	Belarus	Severe	3
	Belgium	Moderate	1
	Hungary	Minimal	1
	Latvia	Severe	1
	Poland	Minimal	2
	Slovakia	Minimal	3
	Slovenia	Moderate	2
	Spain	Critical	2
	Ukraine	Severe	1
Oceania	Australia	Critical	3
	New Zealand	Severe	2

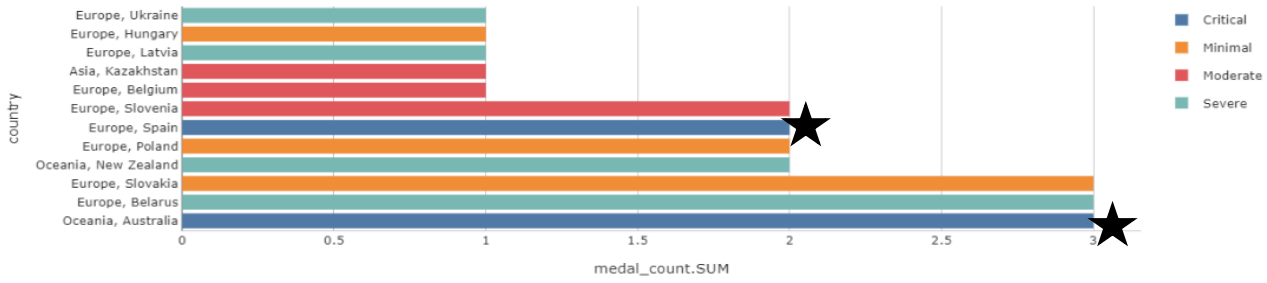
Interesting Trends:

Australia & Spain – Critical

Ukraine, Latvia, New Zealand, Belarus - Severe

session.widget

New stacked bars



Query 9: Could you provide an analysis on how life expectancy correlates with Olympic performance across different regions?

region	life_expectancy	medal_count.SUM
Africa	High	15
	Low	47
	Medium-High	56
	Medium-Low	136
Asia	High	756
	Medium-High	384
	Medium-Low	37
Europe	High	3,303
	Medium-High	554
North America	High	1,457
	Medium-High	43
Oceania	High	370
	Medium-High	2
South America	High	117
	Medium-High	46

Correlation between Life Expectancy and Performance:

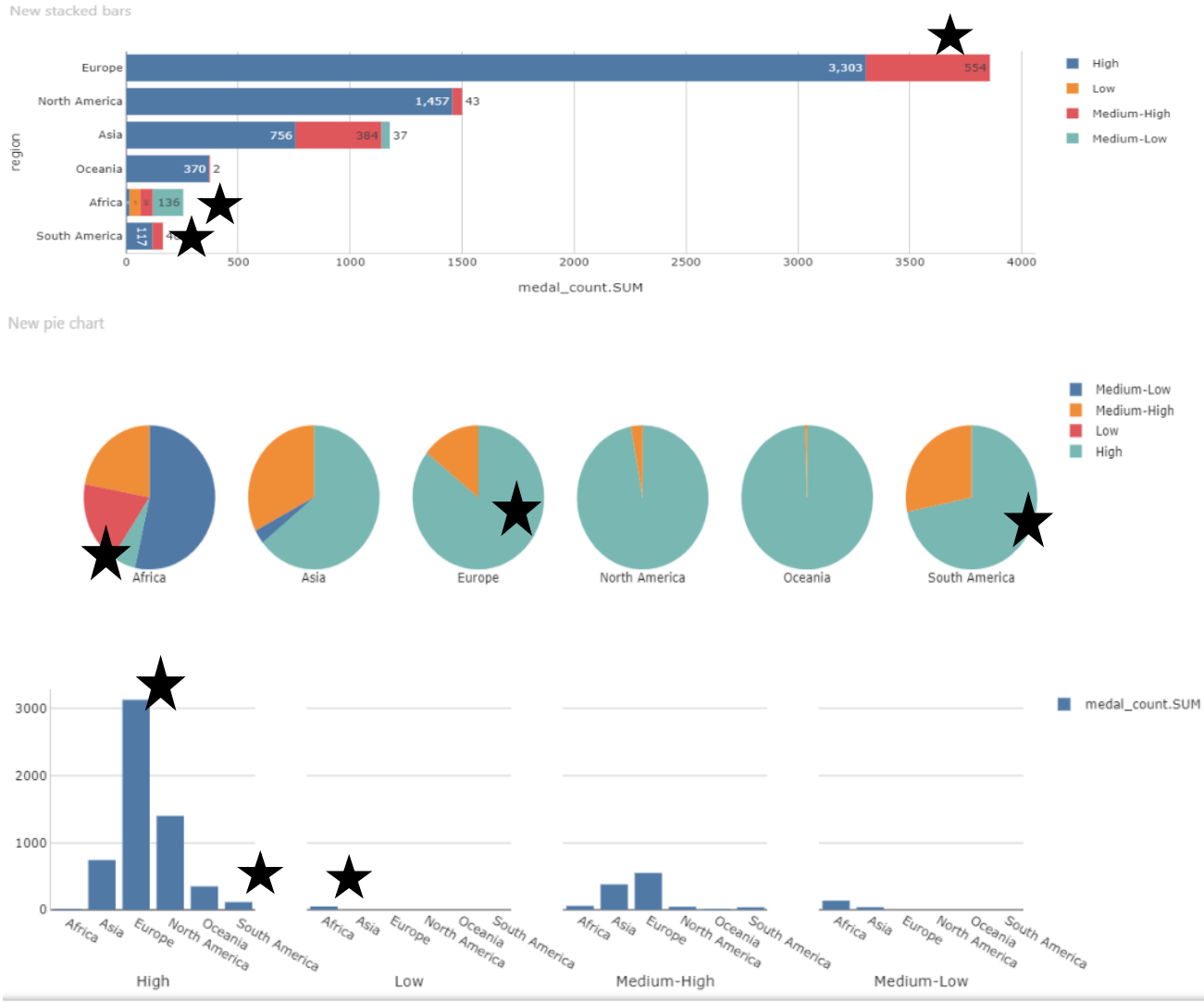
Africa – Low Performance and low life expectancy

Europe – High Performance and low high expectancy

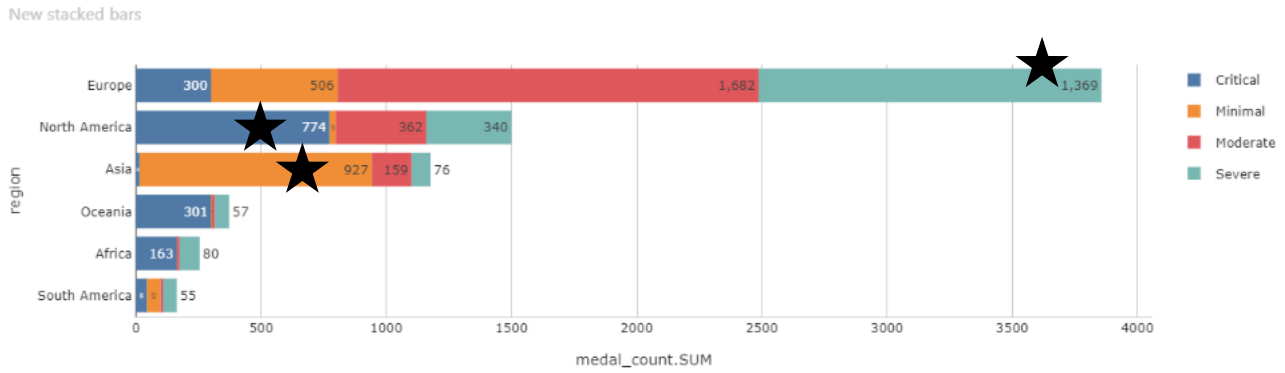
South America (Outlier) – Low Performance, still high expectancy, needs further investigation.

Conclusion = Life expectancy is going hand in hand with performance, with some exceptions.

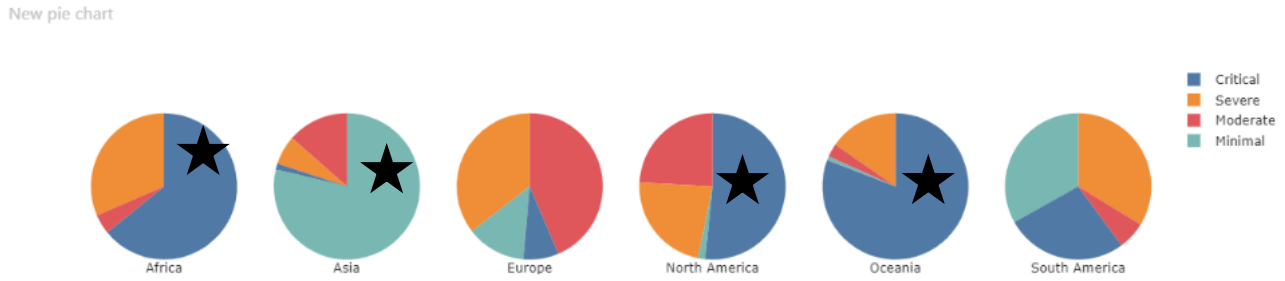
```
session.widget #Life Expectancy
```



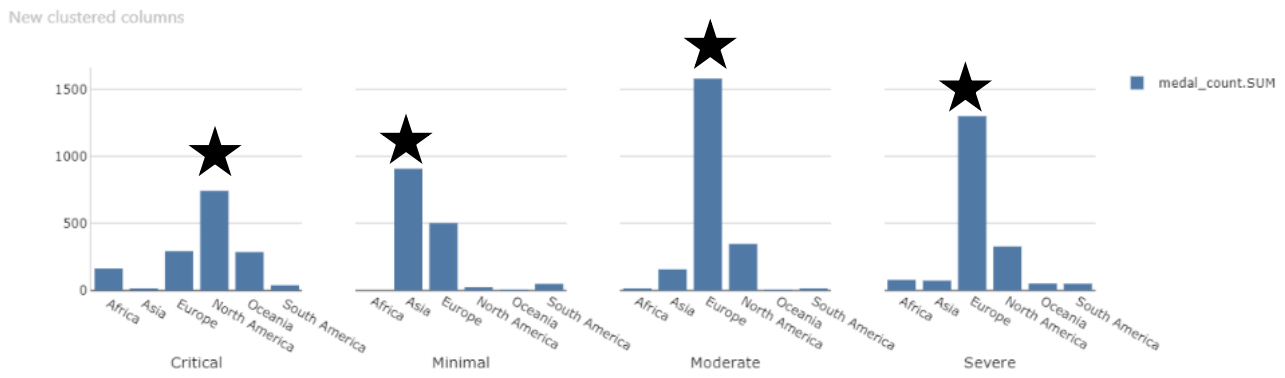
Query 10: Could you provide an analysis on how impact of depression levels correlates with Olympic performance across different regions?




```
session.widget #Depression Levels
```




```
session.widget #Depression Levels
```



<u>Correlation between Depression Levels and Performance:</u>

North America – 2nd best performer and highest number of critical levels

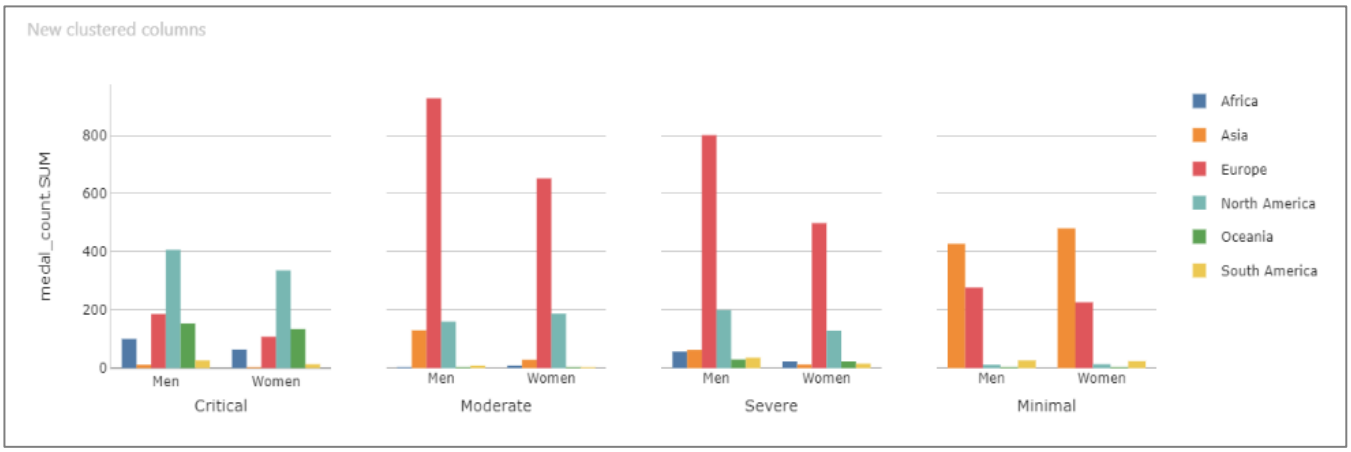
Europe – Best performer and highest number of severe levels

Asia – 3rd best performer and highest number of minimal levels

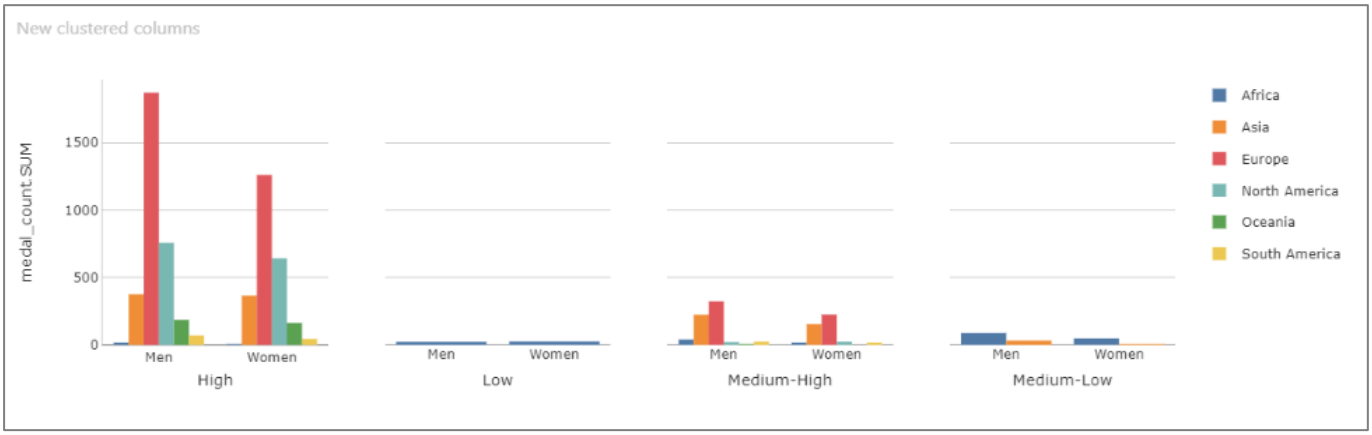
Africa – Low performance, still large number of critical levels

Conclusion = High performing countries suffer from more critical and severe depression levels, needs investigation into reasons, though Africa being low performer also has critical depression levels

Query 11: Could you provide a detailed analysis of the impact levels of depression categorized by gender and region among athletes?



Query 12: Could you provide a detailed analysis of the life expectancy of athletes categorized by gender and region among athletes



Association Rule Mining

Dataset: For association rule mining, I have chosen Olympic medals dataset. Each row in this dataset corresponds to an individual athlete, providing a connection between the athlete and their performance in terms of the medal won. The dataset also ties athletes to their respective countries and provides links to their Olympic profiles, which could be rich in further detail.

Reasons for choosing Olympic Medals Dataset:

- **Pattern Identification**: The dataset can help in identifying patterns of winning, such as whether certain countries consistently perform well in mixed doubles curling events.
- **Performance Analysis**: By mining association rules, one could potentially uncover correlations between athlete participation and medal attainment, possibly leading to insights on the success factors in curling.
- **Strategic Decisions**: Sports organizations, coaches, or even country Olympic committees might use insights from such an analysis to make strategic decisions regarding team compositions, training investments, and talent identification.

Steps followed for Association Rule Mining:

1. **Setting the Minimum Support and Confidence:** Let us assume that the minimum threshold for support is 0.3 and confidence is 0.5 that the rules must meet.
2. **Generating Frequent Itemsets with minimum support**: Using Apriori algorithm to find all itemsets in the dataset that meet the minimum support threshold. Here the length of itemset is being considered as 2 (since k>=1) .

```
#Finding the frequent itemsets
frequent_itemsets = apriori(arm_df,min_support=0.2,use_colnames =True)

#Check the length of rules
frequent_itemsets['length']=frequent_itemsets['itemsets'].apply(lambda x: len(x))

#Assume the length is 2 and the min support is >= 0.3
frequent_itemsets[ (frequent_itemsets['length']==2) &
                   (frequent_itemsets['support']>=0.3)]
```

	support	itemsets	length
9	0.471033	(Men, Athlete)	2

So here we have figured out that

- Frequent itemset (Men, Athlete), support is 0.47 > min support = 0.2
- Antecedents = Men
- Consequents = Athlete

3. **Generate Rules from Frequent Item sets with minimum confidence = 0.5**

```
#Assuming the min confidence is 0.5
rules_con = association_rules(frequent_itemsets, metric="confidence",min_threshold=0.5)
rules_con
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(BRONZE)	(Athlete)	0.347006	0.696548	0.246071	0.709125	1.018056	0.004364	1.043238	0.027161
1	(GOLD)	(Athlete)	0.327649	0.696548	0.225699	0.688845	0.988941	-0.002524	0.975245	-0.016359
2	(Men)	(Athlete)	0.642116	0.696548	0.471033	0.733563	1.053141	0.023768	1.138926	0.140994
3	(Athlete)	(Men)	0.696548	0.642116	0.471033	0.676239	1.053141	0.023768	1.105394	0.166284
4	(SILVER)	(Athlete)	0.325345	0.696548	0.224778	0.690891	0.991879	-0.001840	0.981699	-0.011991
5	(Women)	(Athlete)	0.291423	0.696548	0.210859	0.723549	1.038764	0.007869	1.097670	0.052665
6	(BRONZE)	(Men)	0.347006	0.642116	0.224639	0.647364	1.008172	0.001821	1.014880	0.012413
7	(GOLD)	(Men)	0.327649	0.642116	0.209246	0.638627	0.994566	-0.001143	0.990344	-0.008061
8	(SILVER)	(Men)	0.325345	0.642116	0.208232	0.640034	0.996757	-0.000678	0.994215	-0.004799

4. **Evaluating the Rules using metrics Lift (minimum lift = 1)**

```
#Assuming the min lift is 1
rules_lift = association_rules(frequent_itemsets, metric="lift",min_threshold=1)
rules_lift
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(BRONZE)	(Athlete)	0.347006	0.696548	0.246071	0.709125	1.018056	0.004364	1.043238	0.027161
1	(Athlete)	(BRONZE)	0.696548	0.347006	0.246071	0.353272	1.018056	0.004364	1.009688	0.058446
2	(Men)	(Athlete)	0.642116	0.696548	0.471033	0.733563	1.053141	0.023768	1.138926	0.140994
3	(Athlete)	(Men)	0.696548	0.642116	0.471033	0.676239	1.053141	0.023768	1.105394	0.166284
4	(Women)	(Athlete)	0.291423	0.696548	0.210859	0.723549	1.038764	0.007869	1.097670	0.052665
5	(Athlete)	(Women)	0.696548	0.291423	0.210859	0.302720	1.038764	0.007869	1.016201	0.122977
6	(Men)	(BRONZE)	0.642116	0.347006	0.224639	0.349842	1.008172	0.001821	1.004361	0.022648
7	(BRONZE)	(Men)	0.347006	0.642116	0.224639	0.647364	1.008172	0.001821	1.014880	0.012413

5. **Filtering the above output for min confidence = 0.5**

```
#Based on min confidence (=0.5),
#output antecedents, consequents, support, confidence and lift.
result_arm = rules_con[['antecedents','consequents','support','confidence','lift']]
result_arm
```

	antecedents	consequents	support	confidence	lift
0	(BRONZE)	(Athlete)	0.246071	0.709125	1.018056
1	(GOLD)	(Athlete)	0.225699	0.688845	0.988941
2	(Men)	(Athlete)	0.471033	0.733563	1.053141
3	(Athlete)	(Men)	0.471033	0.676239	1.053141
4	(SILVER)	(Athlete)	0.224778	0.690891	0.991879
5	(Women)	(Athlete)	0.210859	0.723549	1.038764
6	(BRONZE)	(Men)	0.224639	0.647364	1.008172
7	(GOLD)	(Men)	0.209246	0.638627	0.994566
8	(SILVER)	(Men)	0.208232	0.640034	0.996757

6. **Resetting the threshold for confidence = 0.7 to obtain top k rules (k>=1)**

```
#Finding the rules whose confidence >= 0.7
new_result_arm = result_arm[result_arm['confidence']>=0.7]
new_result_arm
```

	antecedents	consequents	support	confidence	lift
0	(BRONZE)	(Athlete)	0.246071	0.709125	1.018056
2	(Men)	(Athlete)	0.471033	0.733563	1.053141
5	(Women)	(Athlete)	0.210859	0.723549	1.038764

7. Rule Interpretation

A. <u>Rule 0: (BRONZE) → (Athlete)</u>

- **Support = 0.246071**: This value indicates that 24.6071% of all transactions in the dataset contain both the antecedent (BRONZE) and the consequent (Athlete). This is a measure of how frequently these two items occur together.
- **Confidence = 0.709125**: There is a 70.9125% probability that transactions containing BRONZE also contain Athlete. This shows how often the rule has been found to be true.
- **Lift = 1.018056**: The occurrence of BRONZE and Athlete together is 1.018 times more likely than it would be if they were statistically independent. A lift value close to 1 suggests a weak association, meaning BRONZE does not strongly influence the occurrence of Athlete.

B. <u>Rule 2: (Men) → (Athlete)</u>

- **Support = 0.471033**: About 47.1033% of the transactions include both Men and Athlete. This higher support indicates that this pairing is more common in the dataset compared to the first rule.

- **Confidence = 0.733563**: There is a 73.3563% chance that transactions containing Men also include Athlete. This rule is slightly more reliable than the first in predicting the presence of Athlete.
- **Lift = 1.053141**: The presence of Men increases the likelihood of also finding Athlete by a factor of 1.053 compared to if they were independent. This suggests a slightly positive association, although it is still relatively weak.

8. Overall Analysis

These rules seem to be exploring the relationship between demographic group men and recipients of the BRONZE medal and their association with being athletes. The rules suggest that:

- Athletes are commonly associated with these demographic groups in the dataset.
- The lift values across all rules are close to 1, indicating only a slight increase in likelihood beyond what would be expected if there were no association at all.

9. Suggestions to Clients

- For Men: Launch sports gear and training programs aimed at male athletes, highlighting attributes or benefits that align with their preferences and buying behaviours.
- For athletes like those who have won Bronze, who may be in the process of training to improve their performance, offer recovery-focused products like muscle relaxants, foam rollers, and personalized dietary supplements.
- Organize exclusive webinars, meet-and-greets with renowned athletes, or members-only competitions. These events can be particularly appealing to sports enthusiasts and could help in building a community around the brand.

10. Conclusion

Each of these suggestions is designed to leverage the findings from our association rule mining to enhance commerce strategies, specifically tailored to the segments identified as significant in the dataset. By focusing on sports gear for men, customized products, and loyalty programs, we can not only increase sales but also strengthen customer relationships and brand reputation in the competitive sports and athletics market.

REFERENCES

Wilson, R., Groen, R., Yambasu, S., Kamara, T., Kushner, A., Remick, K., & Masuoka, P. (2015). Geographic information systems (GIS) in global public health: A Sierra Leone case study. Annals of Global Health, 81(1), 83. https://doi.org/10.1016/j.aogh.2015.02.695

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley.

McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.

Harinarayan, V., Rajaraman, A., & Ullman, J. D. (1996). Implementing data cubes efficiently. Proceedings of the 1996 ACM SIGMOD international conference on Management of data, 205-216. https://doi.org/10.1145/233269.233333

Chaudhuri, S., & Dayal, U. (1997). An Overview of Data Warehousing and OLAP Technology. ACM Sigmod Record, 26(1), 65-74. https://doi.org/10.1145/248603.248616

The Pandas development team. (n.d.). Pandas Documentation: Dataframe and Series. Retrieved April 19, 2024, from https://pandas.pydata.org/pandas-docs/stable/index.html

Atoti. (n.d.). Atoti: Agile business analytics. Retrieved April 19, 2024, from https://www.atoti.io

OpenAI. (2023). ChatGPT: Optimizing language models for dialogue. Retrieved April 19, 2024, from https://www.openai.com/chatgpt

Data Sources:

Olympic Summer & Winter Games. (2022). Olympic hosts.csv and olympic medals.csv data files covering the Olympic Games from 1896 to 2022.

Our World in Data. (n.d.). Mental-illness.csv and life-expectancy.csv data files. The datasets include information on Disability-adjusted life years (DALYs). Retrieved April 10, 2024, from https://ourworldindata.org

International Monetary Fund. (n.d.). Global Population.csv. Retrieved April 10, 2024, from https://www.imf.org

World Bank. (n.d.). Economic data.csv from world-development-indicators. Retrieved April 10, 2024, from https://data.worldbank.org/indicator

World Population Review. (2024). List of countries_areas-by-continent-2024.csv. Retrieved April 10, 2024, from https://worldpopulationreview.com