# Introduction

The Olympic torch was lit in April 1896 in Athens, Greece for the very first time, which marking the moment that this huge international sports event, combining art and sport, was introduced to the public. Every four years, athletes from all over the world were prepared and joined for this game, showcasing not only their amazing sports skills to the audience but also their will to represent their nations with pride. In addition, hosting such huge events is beneficial to the host country in many ways. For instance, the Sydney Olympics helped Australia's economy, resulting in a staggering 6.1 billion Australian dollars more in GDP. Investigating the profound effects of the Olympics on public and organizations is becoming increasingly important, as is developing new tools for analyzing data and drawing conclusions about these matters.

Regardless of whether country hosted the Olympics or won any gold medals, the Olympic have a significant influence on countries all around the world. As a result, gathering information on the Olympics and its profound effects on society and institutions, as well as developing new tools for data analysis and conclusion drawing, are becoming more and more important.

The General Administration of Sport of China is responsible for overseeing and regulating sports activities, which makes it the powerhouse of the Chinese sports development. It attempts to encourage public engagement in physical fitness activities while enhancing national sports capabilities and competitive ability, particularly in worldwide sports competition. The establishment of a data warehouse will provide the administration with a wealth of Olympic-related information, such as host nations, past results, and socioeconomic backgrounds, to help with these endeavors. With the use of this technology, the government will be able to customize investments and policies to increase public engagement in sports and support for Olympic athletes. These decisions will be based on a thorough examination of both economic and global sports trends.

Another client is a potent national public health agency in the United States called The Centers for Disease Control and Prevention (CDC), The primary goal of the CDC is to prevent and investigate disease, trauma, and disability to preserve the general public's health and well-being. It is beneficial to gain a deeper understanding of the relationships between many aspects, including mental health, medical standards, and the economy in different nations for CDC by evaluating and developing relationships among birth rates, sports levels, and health expenses created in a data warehouse.

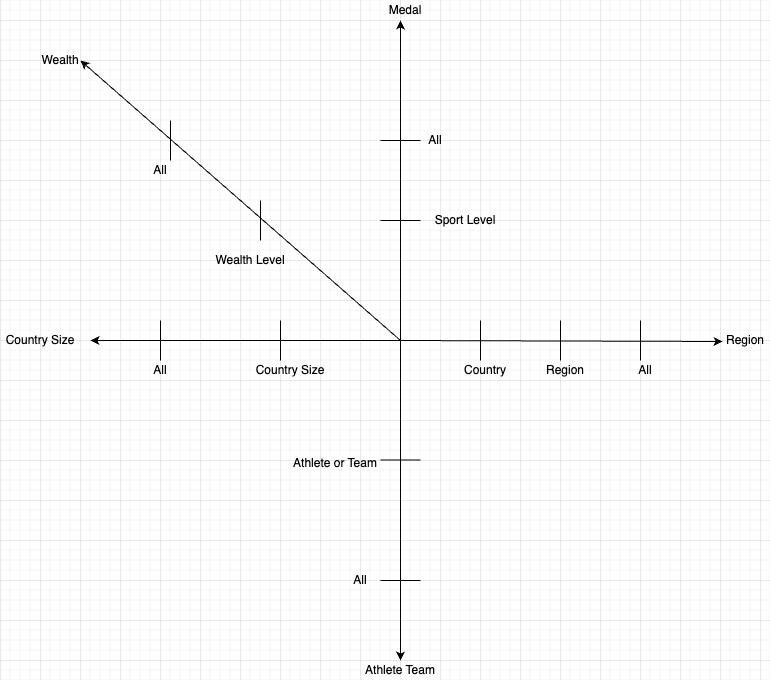
Using seven comma-separated value (CSV) files, this project aims to construct a data warehouse that will be utilized to simulate and analyze three business queries. To provide a thorough overview, this project is divided into eight parts. Firstly, the project assumes three hypothetical clients who might benefit from the data. The subsequent section will provide summaries and relevant information about these CSV files. The related Star Schema diagrams for these three business queries are introduced in the third section. The fourth session, which is the most important part of this project, includes the process explanation of the Fact table and Dimension table, and data warehouse modeling. Introduction to the process of ETL stages and data cleaning are provided as part of the fifth step. OLAP cube design will be explained in the sixth part. After that, data visualization for these business queries will be shown in the seventh part. Finally, the project will explain some questions regarding these queries.

Firstly, we need to explore the Olympic dataset and create conceptual hierarchy for each dimension. In StarNet, each footprint represents a level of the dimention, culminating in an "All" at the top. Taking "region" as an example, its conceptual hierarchy is divided into three levels: total, continent, and country. These three levels are reflected in the StarNet model

许多不同颜色的地图

中度可信度描述已自动生成

In the StarNet model below, we can see that the chart is designed to analyze the relationship among the sports level of various countries, country size, wealth, and country. Each line in the chart represents these factors, with footprint on the lines indicating different levels of detail within the model. This means that the generality or range of data changes with the level of abstraction. The variation in detail depends on the OLAP process applied to the model. For example, in the "Region" dimension, users can generalize data from the country to the continent level through roll up, or access more detailed information from the state to the national level through roll down.



The StarNet for Client 2 below focuses on health and costs. The Starnet is connecting dimensions such as Country Size, Region, Country, Sports Level, Internet Level, and Host. These dimensions highlight the importance of analyzing health costs based on geographical, technological, and sports-related factors to understand how these factors affect medical expenditures or outcomes in different regions. The chart also suggests a relationship between the level of internet access and healthcare costs.

图表, 散点图

描述已自动生成

Business Queries and StarNet

Five business queries are created for each client, and the queries are listed as follows,

Client 1: The General Administration of Sport of China

1. How many times have countries with huge population host in total?

图表

描述已自动生成

1. How many medals did the wealthiest continent in 2020 win at the Tokyo Olympics?

图表, 雷达图

描述已自动生成

1. What was the GDP growth rate of the country that won the most medals in 2020?

图片包含 图表

描述已自动生成

1. Do populous countries prefer individual or team sports?

图片包含 图示

描述已自动生成

1. How many medals has the country that aces in individual sports won in their favorite sport?

图示

中度可信度描述已自动生成

Client 2: The Centers for Disease Control and Prevention

1. How much more does the continent with the strongest sports power spend on health compared to the continent with average sports power?

图表, 雷达图

描述已自动生成

1. How much more does the continent with the strongest sports power spend on health compared to the continent with average sports power?

许多的地图

描述已自动生成

1. Do governments with middle health costs prefer to pay for their citizens' healthcare themselves, or do they let their citizens bear the costs?
2. How does the GDP look like for middle countries with strong internet systems after hosting the Olympics?
3. The change in the number depress disorder in all continents that not good at sports, specifically before and after the 2002 Olympics.

Star Schema

The databases of two clients, that are shown below, are both constructed by the star schema. For the first client, there are five dimension tables, including countrysize, region, sports level, economy level, and the size of gamers. The measuer used on the fact table the number of medals from the 2021 Olympics, GDP, GDP growth rate, the number of medals in sports each country good at, the number of times hosting the Olympics, and the average number of medals in individual and team competitions.

A screenshot of a computer

Description automatically generated  
In the second client, there are also five dimension tables surrounding the fact table, including health expenditure level, internet level, country/region, sports capability, and whether the Olympics have been hosted or not. On the other hand, the fact table contains the measures including the average yearly GDP growth rate, the change amount of depression disorder in 2002, GDP per capita, and the difference of health expenditures between 2002.

A screenshot of a computer

Description automatically generated

Extract, Transform, and Load

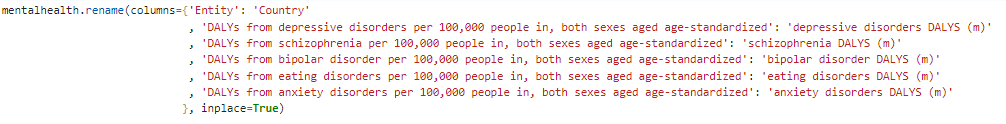
We can see that, from the Olympic dataset, there are a lot of data issues and inconsistencies, such as missing values. Therefore, I used python, Microsoft VS code, and sql to extract, transform and load the cleaned and appropriate data, then using Jupyter notebook to create and populate the database in the pgadmin4 server.

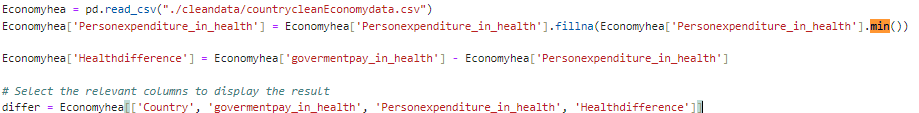
The first step of the ETL process is to explore and analysis all the relevant data we are asked to answer the business queries. Based on the queries analysis above, we have to remodel and structure the data in an appropriate schema in order to extract the necessary information.

A screenshot of a computer

Description automatically generated

Firstly, as the image shown above, I used the read\_csv() statement from the pandas tookit to import all data into Python and assign them to different variables, also we will observe their data structure. These datasets include both numerical and textual data, which is a big advantage for our data analysis. However, before proceeding with the analysis, it is necessary to rename each column with long column names by using .rename() to benefit following analysis tasks.



To make sure the completeness and rationality of the data, it is essential to clean rows and columns and deal with missing values. In the Olympic data package, for example, all data contain a certain amount of missing values. Taking "Economic data.csv" as an example, we use the .isna().sum() statement to check the number of null values in each column. We find that most of the countries with missing values come from either poor countries or countries in wartime. Therefore, using the average value of the entire column is not a good choice to fill the blanks Instead, we choose to fill in the missing values with the minimum value from the list, selecting the most appropriate data without affecting the analysis.

In the Olympic data, certain values for countries and regions need to be removed from the data frame, such as G7, Asia, and middle income. The data from these entities have a minimal impact on the Olympic games and are not helpful for our research at the country level. Storing data in this manner introduces a level of inequality. For instance, D7 and groups of low-income countries do not specify which countries they include or how to distinguish the standard for low income, leading to confusion

Fixing country names is also a challenge in the ETL process. In all six files in the Olympic dataset, each table has its own country column, and the names are different to each other. To facilitate data analysis in future, we will use the country column in list-of-countries\_areas-by-continent-2024.csv as the standard to modify other tables. For other tables, we will remove the names that are not considered as country, and the country name that no longer exist due to historical and political reasons. After that, for cases where the same country that has different names in different tables, we will use the .difference() statement as below to filter out incorrect names, and then correct and modify the names in VS Code.A screen shot of a computer code

Description automatically generated

Next, we will proceed with deeper processing and transformation of the data. We will create a new column in the fact table to store calculated values based on population numbers. The detailed process is as follows: We use an if statement to go through the population data of each country and then compare each piece of data against three defined levels. The number 1 represents a population size of more than 40 million, 3 represents a population of less than 2 million, and for the rest of the population sizes, we use 2 to represent them. These numbers 1, 2, and 3 correspond to the classifications in DimCountrySize.

A white background with colorful text

Description automatically generated

In the fact tables of both clients, there are many examples like this, such as DimWealth. To categorize countries by different economic levels, we use If statement to go through the GDP of all countries, and then divide the highest GDP value by three. This allows us to establish three levels. Countries are then classified using the numbers 1, 2, and 3 according to this classificationA computer screen shot of a code

Description automatically generated  
For the measures of medals in the fact table and the dimension table, although the measures we need and the data corresponding to the dimension table in the olympic\_medal.csv are in text format, we can still process and convert this data into numeric information. For example, in the fact table, if we need to calculate the total number of medals for each country, we can use the .groupby() statement to group by country, and then count the number of medals for each country. This processing not only transforms text data into a computable numeric form but also benefits future data analysis and decision-making.

A computer code with colorful text

Description automatically generated

1.5 Implementation

1.5.1 Create table schema

After cleaning the Olympic data, my next step is to create database, tables, and their schema in pgAdmin 4. This process will be constructed through Python script, to ensure the uniqueness of the database and tables with no other database and tables with the same names. After that, we will create a database named ‘Project1’ and define the schemas of fact tables and dimension tables, including their primary and foreign key relationships. This is to follow the star schema data warehouse design, optimizing data analysis and query efficiency.

A screen shot of a computer program

Description automatically generated

Next, I will use the SQL “CREATE TABLE” statement to construct the schema of mentioned tables above. Let the dimension table “dimRegion” as an example, I will define the column named “countryID” as the primary key to uniquely identify each country. The table will include other necessary columns, and each column clearly specify the data type, whether null values are allowed, and other constraints to ensure data integrity and accuracy.

A screenshot of a computer screen

Description automatically generated

Finally, each dimension table will be constructed with a primary key to follow the star schema, and the fact table will link to them through foreign keys. This step ensures that the fact table and dimension tables can be connected efficiently.

A screenshot of a computer code

Description automatically generated

1.5.2 Upload the data

Completing the construction of the table schema in pgadmin 4, we will use Python scripts to import the cleaned data from .csv files into the dimension and fact tables in pgAdmin. This process refers to the code from Lab 3. By using the ‘read\_csv’ statement, we can read cleaned data from csv files and insert them into the schema The directory path where the CSV files are stored is set to ./projectdata/Client1 or ./projectdata/Client2.

2.0 cube and virtualization

2.1 Cube

After successfully connecting to pgAdmin using python and creating two fact tables and dimensional tables for two clients, we new continue to use Python to line the data to pgAdmin and create data cubes for both clients. Two following plots are the data relationship diagrams between the fact table dimension tables. Th appearance of data cube enhance the possibilities and operability of data analysis and organization, which makes it better than other sql database.

This is the data relationship diagram for Client 1.A diagram of a company

Description automatically generated with medium confidence

This is the data relationship diagram for Client 2A screenshot of a computer

Description automatically generated

2.2 Data Cube Hierarchy

Data cubes play a significant role in data warehouse by creating meaningful concept hierarchies to establish relationships among data. Consider the concept hierarchy for Client 1 below as an example, after we establish a hierarchical structure between continents and countries, we can use OLAP operations such as drill-down to obtain detailed information about each country in a continent, or to explore which countries are associated with specific continent. Once the hierarchical structure of countries is successfully constructed, we can delve into the relationships between features by performing roll-up and drill-down operations.

A screenshot of a computer program

Description automatically generated

2.3 Virtualization

Business Queries

After finishing the data preprocessing mentioned above, I constructed OLAP operations, including slice, roll down, and roll up, to clean hierarchies and measures. Through these processes, I obtained cleaned and organized data, allowing me to effectively answer the queries posed by two clients.

Client 1

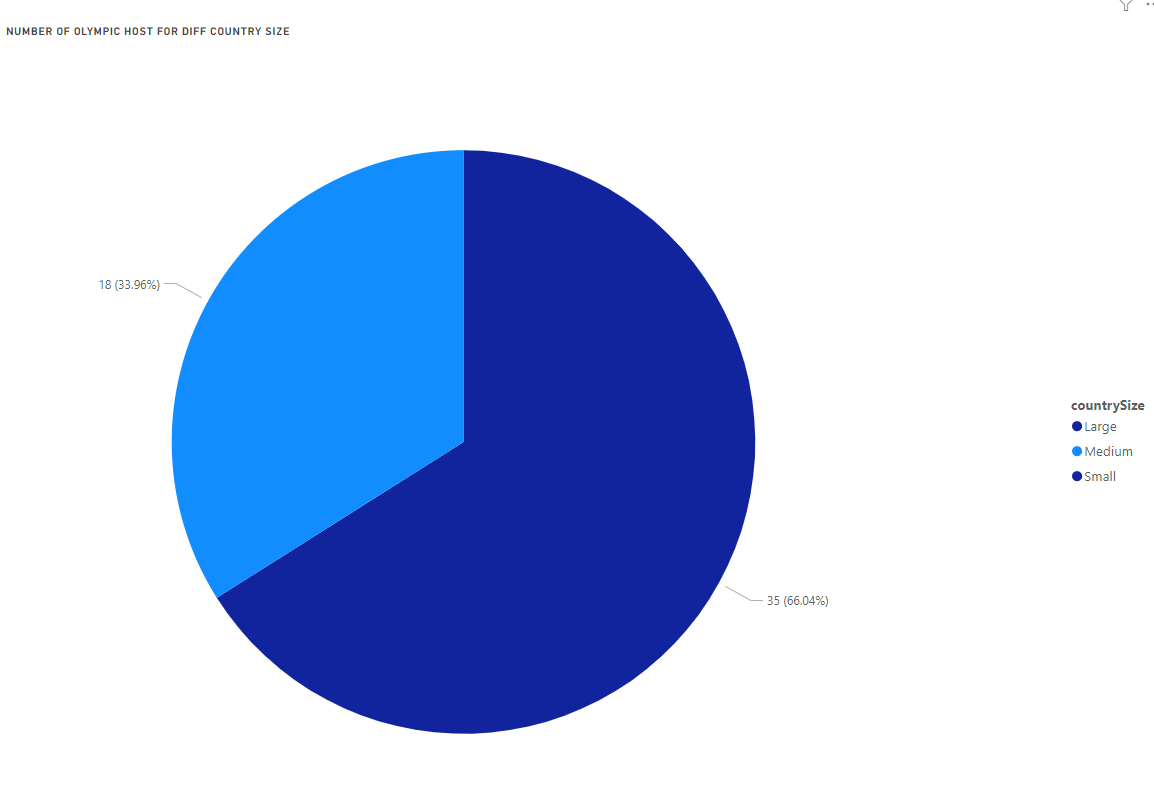
1. How many times have countries with huge population hosted in total?

In this query, I selected the dimensions of countrysize and the measure of the number of hold, with the following data:

A screenshot of a computer

Description automatically generated

And the plot is as follows



1. How many medals did the wealthiest continent in 2020 win at the Tokyo Olympics?

In this query, I selected the dimensions of continent and the measure of the number of ‘2021 gode medel’ and GDP per capita with the following data:

A screenshot of a black and white screen

Description automatically generated

And the plot is as followsA graph with blue and yellow bars

Description automatically generated

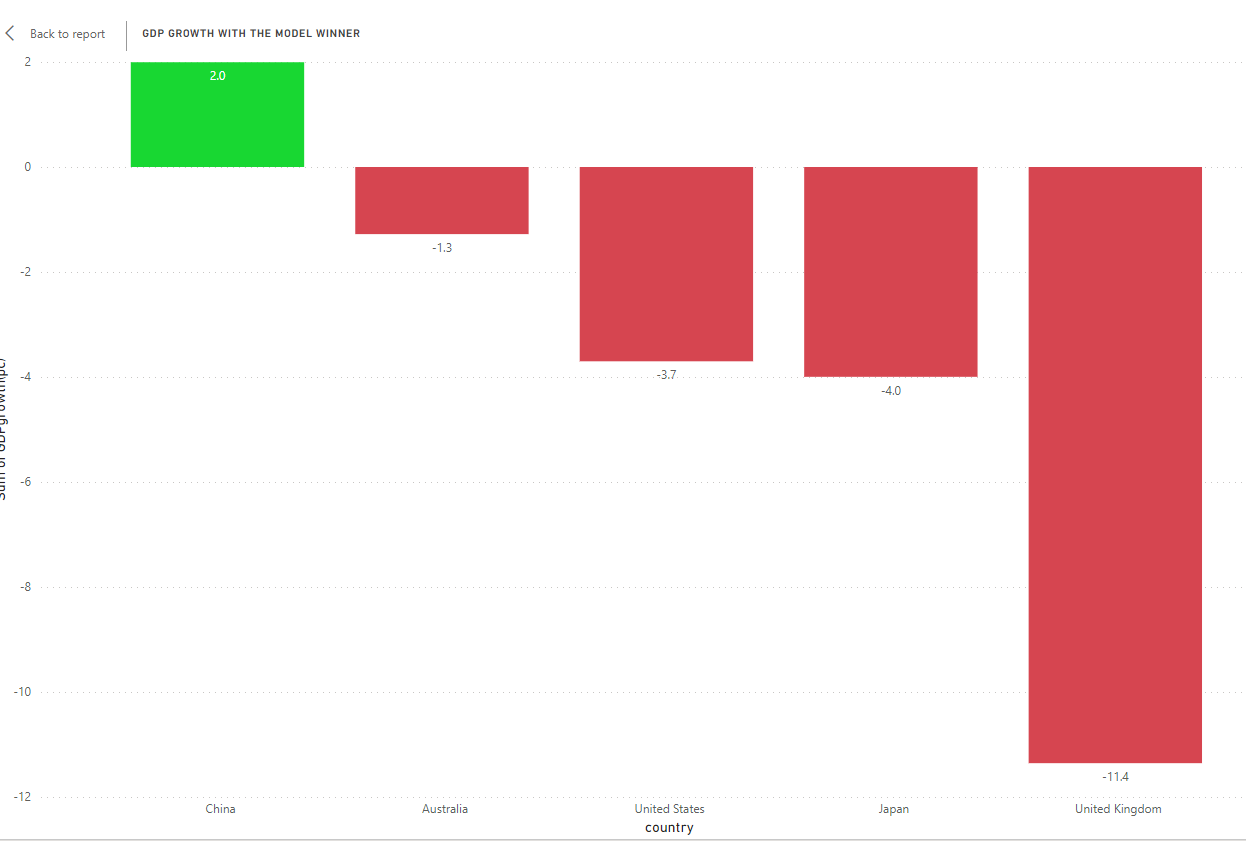
1. What was the GDP growth rate of the country that won the most medals in 2020?

In this query, I selected the dimensions of continent and the measure of the number of ‘2021 gode medel’ and GDP growth per capita with the following data:

A screenshot of a black screen

Description automatically generated

And the plot is as follows,



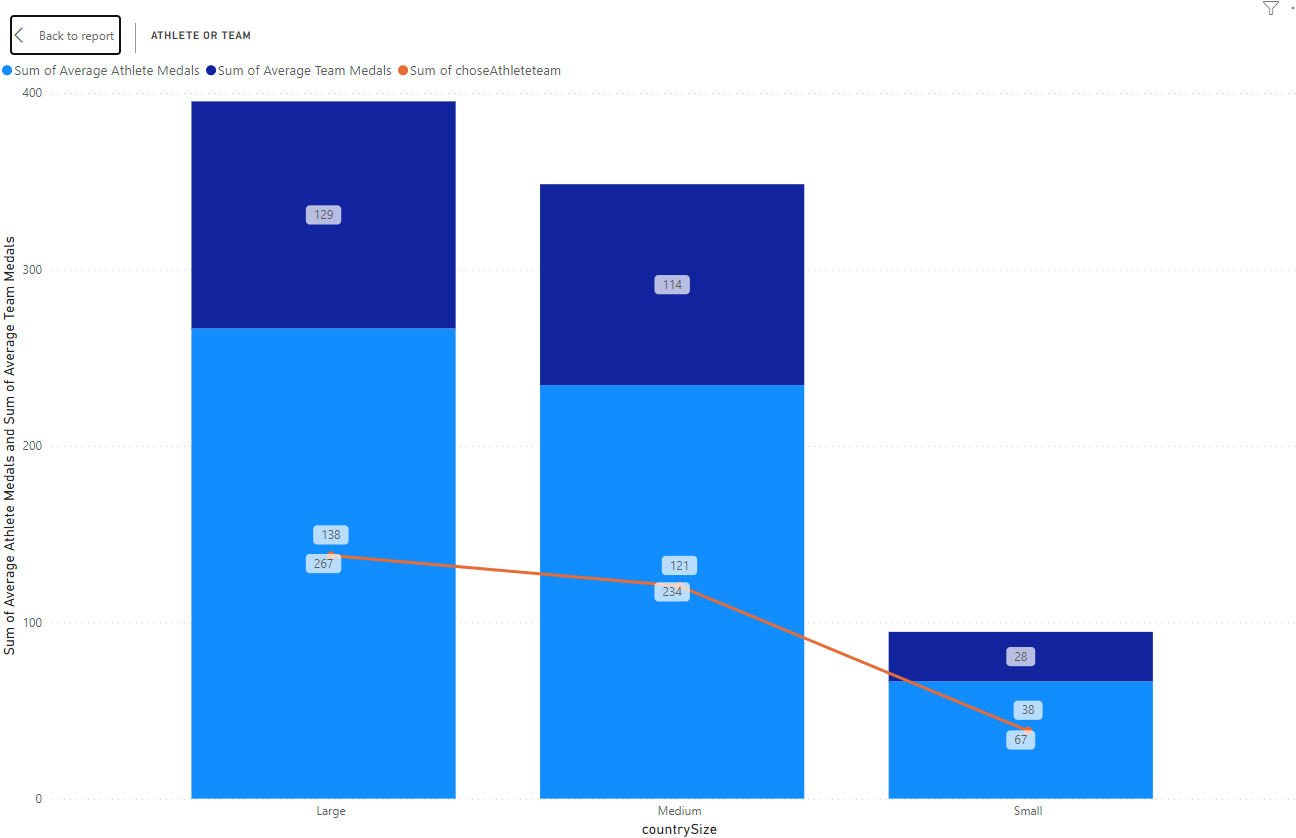
1. Do populous countries prefer individual or team sports?

In this query, I selected the dimensions of countrysize and the measure of Average Athlete Medals and Average Team Medals with the following data:

A screenshot of a black screen

Description automatically generated

And the plot is as follows,



1. How many medals has the country that aces in individual sports won in their favorite sport?

In this query, I selected the dimensions of continent and sportlevel and the measure of medel for strongest sport with the following data:

A screenshot of a black and white screen

Description automatically generated  
and the plot is as follows

A colorful circle with text

Description automatically generated

Client 2

1. Which countries have passed through a decline in their year-over-year population growth rate after the beginning of the 21st century?

In this query, I selected the dimensions of region and the measure of Average Annual Growth Rate (%) with the following data:

A screenshot of a graph

Description automatically generated

And the plot is as follows

A yellow line with black text

Description automatically generated with medium confidence

1. How much more does the continent with strongest sports power spend on health compared to the continent with average sports power?

In this query, I selected the dimensions of region and sportlevel in GDP and the measure of health expenditure(GPT%) and GDP per capita with the following data:

A screenshot of a black and white screen

Description automatically generated

And the plot is as follows

A blue circle with a few blue circles

Description automatically generated

1. Do governments with middle health cost prefer to pay for their citizens' healthcare themselves, or do they let their citizens bear the costs?

In this query, I selected the dimensions of country and health cost in GDP and the measure of health cost difference of Change with the following data:

A screenshot of a computer screen

Description automatically generated

And the plot is as follows

A green and pink bar graph

Description automatically generated

1. How does the GDP look like for middle countries with strong internet systems after hosting the Olympics?

In this query, I selected the dimensions of region, whether host before and Internet level and the measure of GDP per capita of Change with the following data:

A screenshot of a black screen

Description automatically generated

And the plot is as follows

A blue and orange bar graph

Description automatically generated

1. the change in the number depress disorder in all continents that not good at sports, specifically before and after the 2002 Olympics.

In this query, I selected the dimensions of region and sport level and the measure of depressive disorders rate of Change with the following data:

A screenshot of a black and white screen

Description automatically generated

And the plot is as follows

A blue and orange rectangular object

Description automatically generated