JACOBS UNIVERSITY

Data Management Plan

Impact of Covid-19 on the Countries' Performance during the Tokyo Summer Olympics 2020

Tanuja Seervi & Bikiran Choudhury

Course: Introduction to Data Management with Python
Instructor: Dr. Carlos Brandt

Table of Contents

Table of Tables

Table of Figures

1. Introduction

The Tokyo Summer Olympics 2020 was postponed by a year due to the pandemic and was later held from 23 July to 8 August 2021 in Tokyo, Japan. It was organised behind closed doors with no spectators to avoid the risk of any outbreaks. 206 countries participated in a total of 339 events in 33 sports and 93 countries won medals.

However, the Covid-19 pandemic has impacted all the different nations to different degrees. Some countries were severely impacted, whereas some countries managed to control the damage with some success. Training and mental health of the participating athletes were also impacted.

2. Objective

The goal of this project to analysis and explore the effect of the Covid-19 pandemic on the Tokyo Olympics 2020 (held in the year 2021).

Specifically, we want to discover if the performance of the nations (in terms of medals earned) in the Tokyo Olympics had any significant deviation from the last couple of Summer Olympic Games (London Olympics 2012 and Rio Olympics 2016).

3. About the Data

3.1 What information is required ?

To successfully execute and achieve the desired objective, we require mainly two types of information:

- first, the performance of the countries in the Olympic games, and
- second, an estimate of how each country has been impacted by the covid-19 pandemic.

The performance of the countries are measured by the total number of medals won in the Tokyo 2020, Rio 2016 and London 2012 Olympic games.

For estimating the impact of covid-19 pandemic on the countries, we are considering the following metrices:

1. the new and cumulative number of covid-19 cases, new and cumulative number of deaths reported due to covid-19.
2. the total number of persons who have been vaccinated with one dose and number of persons who are fully vaccinated (two doses) against covid-19.

We are also considering the GDP of a country as a proxy for estimating the particular country's access to resources to handle a pandemic.

3.2 Flow of Data

The overall flow of data during this project can be represented as below:

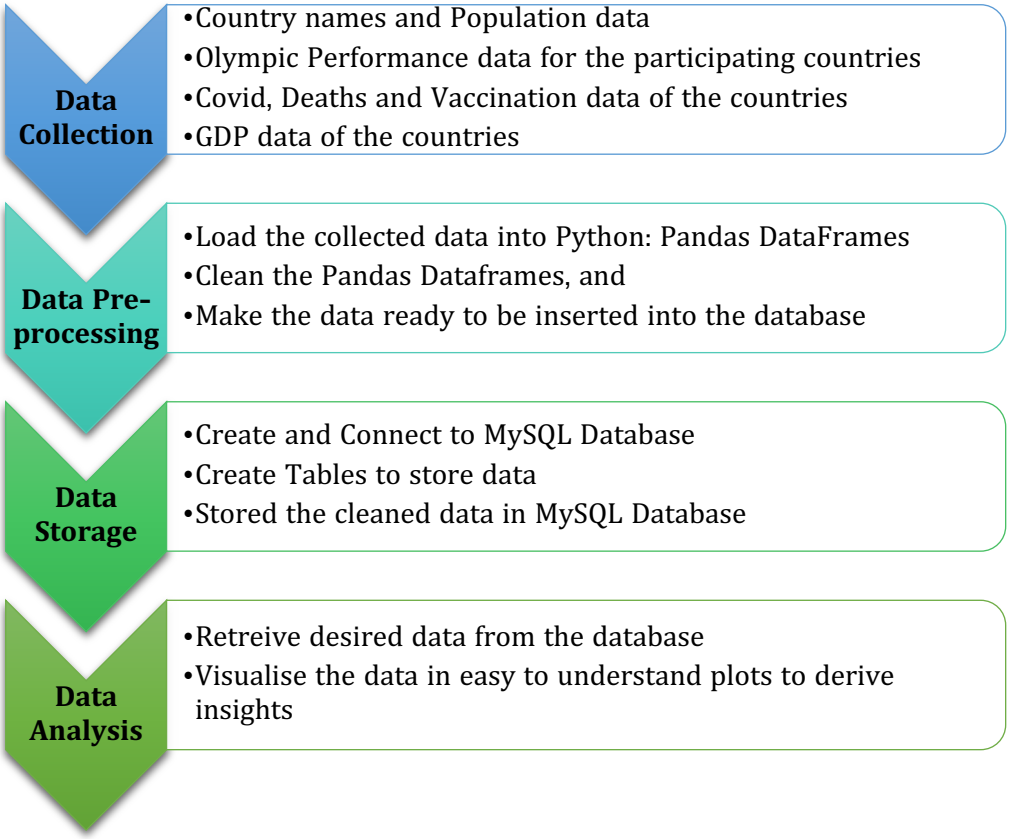

Figure 1: Data Flow Chart

3.3 Description of the data

A detailed description of the various datasets utilised in the project is captured in this section. The datasets mentioned in the below sub-sections are stored as individual database tables in a MySQL database.

3.3.1 Country Name and Population data

These dataset contains the country names and population data of all the countries for the year 2020 and 2021.
And we are using the country names from this dataset as the standard names of countries across all the datasets.

This dataset has the below information:

Column Name	Description	Data Type	Sample Value	Null/ Not Null	Unique / Not Unique
country_name	The name of the country	string	"India"	Not Null	Unique
pop_2020	Total population of the country in the year 2020	integer unsigned	5829120	Not Null	Not Unique
pop_2021	Total population of the country in the year 2021	integer unsigned	5939234	Not Null	Not Unique

Table 1: Country Name and Population

3.3.2 Country-wise Olympic Medals Data

We have three different datasets for the three Olympic games: Tokyo 2020, Rio 2016 and London 2012 for the participating countries.

And each of the datasets have the below information:

Column Name	Description	Data Type	Sample Value	Null/ Not Null	Unique / Not Unique
country_name	The name of the country participating in the Olympic games	string	"India"	Not Null	Unique
gold_medals	Total number of Gold medals won by the country.	integer	5	Not Null	Not Unique
silver_medals	Total number of Silver medals won by the country.	integer	3	Not Null	Not Unique
bronze_medals	Total number of Bronze medals won by the country.	integer	0	Not Null	Not Unique
total_medals	Total medals won by the country.	integer	8	Not Null	Not Unique

Table 2: Country-wise Olympic Medals Counts

3.3.3 Country-wise Covid Cases, Deaths and Vaccinations Data

This dataset holds the date-wise data of the number of new and cumulative cases of covid-19, new and cumulative deaths due to covid-19, and number of people vaccinated with one dose and two doses (fully vaccinated) against covid-19 for the countries.

This dataset has the below information:

Column Name	Description	Data Type	Sample Value	Null/ Not Null	Unique / Not Unique
country_name	The name of the country	string	"India"	Not Null	Unique
date_reported	Date of reporting of the corresponding data	date	2020-04-19	Not Null	Not Unique
cumulative_cases	Total cumulative cases of the country	integer unsigned	12279	Not Null	Not Unique
new_cases	New cases reported on this day	integer	50	Not Null	Not Unique
cumulative_deaths	Total cumulative deaths of the country	integer unsigned	2319	Not Null	Not Unique
new_deaths	New deaths reported on this day	integer	10	Not Null	Not Unique
people_vaccinated	Total number of people vaccinated with atleast one dose	integer	2319	Not Null	Not Unique
people_fully_vaccinated	Total number of people who are fully vaccinated	integer	2319	Not Null	Not Unique

Table 3: Country-wise Covid Cases, Deaths and Vaccinations Data

3.3.4 GDP of the countries

This dataset has the GDP figures of the countries from the year of 2012 to 2021.

This dataset has the below information:

Column Name	Description	Data Type	Sample Value	Null/ Not Null	Unique / Not Unique
country_name	The name of the country	string	“India”	Not Null	Unique
gdp_2012	The total gross domestic product of the country for the year of 2012 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2013	The total gross domestic product of the country for the year of 2013 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2014	The total gross domestic product of the country for the year of 2014 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2015	The total gross domestic product of the country for the year of 2015 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2016	The total gross domestic product of the country for the year of 2016 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2017	The total gross domestic product of the country for the year of 2017 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2018	The total gross domestic product of the country for the year of 2018 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2019	The total gross domestic product of the country for the year of 2019 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2020	The total gross domestic product of the country for the year of 2020 (in Million Dollars)	float	5.4	Not Null	Not Unique
gdp_2021	The total gross domestic product of the country for the year of 2021 (in Million Dollars)	float	5.4	Not Null	Not Unique

Table 4: Country-wise GDP Data

3.4 Sources of Data

The data necessary for this project is collected from the websites of worldwide organisations, which manages and tracks these information. All the data are collected in .csv files.

The respective data sources are:

3.4.1 Country Name and Population:

Dataset	Source Name	Link
Population_2020-21.csv	UN-world 2019 review	https://worldpopulationreview.com/countries

Table 5: Source detail for Country Name and Population

Sample Raw Data:

{“cca2”,”name”,”pop2021”,”pop2020”,”pop2050”,”pop2030”,”pop2019”,”pop2015”,”pop2010”,”pop2000”,”pop1990”,”pop1980”,”pop1970”,”area”,”Density”,”GrowthRate”,”WorldPercentage”,”rank”
“AF”,”Afghanistan”,”39835.4280”,”38928.346”,”64682.974”,”48093.578”,”38041.754”,”34413.603”,”29185.507”,”20779.953”,”12412.308”,”13356.511”,”11173.642”,”652230”,”61.0757”,”1.0233”,”0.0051”,”37”
“AL”,”Albania”,”2872.9330”,”2877.797”,”2424.061”,”2786.974”,”2880.917”,”2890.513”,”2948.023”,”3129.243”,”3286.073”,”2682.690”,”2150.707”,”28748”,”99.9351”,”0.9983”,”0.0004”,”140”
“DZ”,”Algeria”,”44616.6240”,”43851.044”,”60923.386”,”50360.749”,”43053.054”,”39728.025”,”35977.455”,”31042.235”,”25758.869”,”19221.665”,”14464.985”,”2381741”,”18.7328”,”1.0175”,”0.0057”,”34”}

3.4.2 Country-wise Olympic Medals Data:

Dataset	Source Name	Link
Tokyo_Medals_2020.csv	Kaggle	https://www.kaggle.com/berkayalan/2021-olympics-medals-in-tokyo

Rio_Medals_2016.csv	Google Data Studio	https://datastudio.google.com/reporting/0ByFLn8uIUCsJVVVaUTRWbk5HWmM/page/oeV
London_Medals_2012.csv	Olympics	https://olympics.com/en/olympic-games/london-2012/medals Note: we prepared csv filed by copying data into excel sheet from above mentioned site.

Table 6: Source detail for Country-wise Olympic Medals Data

Sample Raw Data:

1. Tokyo_Medals_2020.csv:
   {Country,Gold Medal,Silver Medal,Bronze Medal,Total,Rank By Total
   Argentina,0,1,2,3,60
   Armenia,0,2,2,4,47
   Australia,17,7,22,46,6}

2. Rio_Medals_2016.csv
   { Country,Gold,Silver,Bronze
   Algeria,0,2,0
   Argentina,3,1,0
   Armenia,1,3,0
   Australia,8,11,10 }

3. London_Medals_2012.csv
   { Country,Gold Medal,Silver Medal,Bronze Medal,Total
   Afghanistan,0,0,1,1
   Algeria,1,0,0,1
   Argentina,1,1,2,4
   Armenia,0,1,1,2 }

3.4.3 Country-wise Covid Cases, Deaths and Vaccinations Data:

Dataset	Source Name	Link
Covid_Vaccination_Data.csv	Our world in Data	https://ourworldindata.org/covid-vaccinations

Table 7: Source detail for Country-wise Covid Cases, Deaths and Vaccinations Data

Sample Raw Data:

{iso_code,continent,location,date,total_cases,new_cases,new_cases_smoothed,total_deaths,new_deaths,new_deaths_smoothed,total_cases_per_million,new_cases_per_million,new_cases_smoothed_per_million,total_deaths_per_million,new_deaths_per_million,new_deaths_smoothed_per_million,reproduction_rate,icu_patients,icu_patients_per_million,hosp_patients,hosp_patients_per_million,weekly_icu_admissions,weekly_icu_admissions_per_million,weekly_hosp_admissions,weekly_hosp_admissions_per_million,new_tests,total_tests,total_tests_per_thousand,new_tests_per_thousand,new_tests_smoothed,new_tests_smoothed_per_thousand,positive_rate,tests_per_case,tests_units,total_vaccinations,people_vaccinated,people_fully_vaccinated,total_boosters,new_vaccinations,new_vaccinations_smoothed,total_vaccinations_per_hundred,people_vaccinated_per_hundred,people_fully_vaccinated_per_hundred,total_boosters_per_hundred,new_vaccinations_smoothed_per_million,new_people_vaccinated_smoothed,new_people_vaccinated_smoothed_per_hundred,stringency_index,population,population_density,median_age,aged_65_older,aged_70_older,gdp_per_capita,extreme_poverty,cardiovasc_death_rate,diabetes_prevalence,female_smokers,male_smokers,handwashing_facilities,hospital_beds_per_thousand,life_expectancy,human_development_index,excess_mortality_cumulative_absolute,excess_mortality_cumulative,excess_mortality,excess_mortality_cumulative_per_million
AFG,Asia,Afghanistan,2020-02-24,5.0,5.0,,,,,0.126,0.126,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,8.33,39835428.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83,0.511,,,,
AFG,Asia,Afghanistan,2020-02-25,5.0,0.0,,,,,0.126,0.0,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,8.33,39835428.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83,0.511,,,,
AFG,Asia,Afghanistan,2020-02-26,5.0,0.0,,,,,0.126,0.0,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,8.33,39835428.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83,0.511,,,,
AFG,Asia,Afghanistan,2020-02-27,5.0,0.0,,,,,0.126,0.0,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,8.33,39835428.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83,0.511,,,,}

3.4.4 GDP of the countries

Dataset	Source Name	Link
GDP_Actual_Value.csv	IMF	https://www.imf.org/external/datamapper/NGDPD@WEO/OEMDC/ADVEC

Table 8: Source detail for Country-wise GDP Data

Sample Raw Data:

{ ″GDP, current prices (Billions of U.S.
dollars)″,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019,2020,2021,2022,2023,2024,2025,2026
,,
Afghanistan,no data,4.367,4.553,5.146,6.167,6.925,8.556,10.297,12.066,15.325,17.89,20.293,20.17,20.616,20.057,18.02,18.883,18.401,18.876,20.136,no data,no data,no data,no data,no data,no data
Albania,1.946,2.229,2.296,2.319,2.29,2.339,2.587,2.566,2.53,2.779,2.221,1.333,0.843,1.461,2.361,2.882,3.2,2.259,2.56,3.209,3.483,3.928,4.348,5.611,7.185,8.052,8.896,10.677,12.881,12.044,11.937,12.899,12.324,12.784,13.246,11.389,11.862,13.053,15.147,15.283,14.828,16.77,18.012,18.931,20.026,21.15,22.308}

4. Processing of the Data

Before the above collected data is stored in the MySQL database, we carry out the following processing steps to clean the data and prepare it to store in MySQL database, which is used for the analysis and visualisations later on.

All the below steps are performed using Python Pandas package functions.

1. Drop non-essential columns and rows from the dataset:
    a. For Tokyo dataset, we drop the column: "Rank by Total"
    b. For Population dataset, we keep only the following columns: "name, pop2021, pop2020".
    c. For Covid and Vaccination dataset, we keep only the following columns: "location, date, total_cases, new_cases, total_deaths, new_deaths, people_vaccinated, people_fully_vaccinated"
    d. For Covid and Vaccination dataset, we remove the rows, have following entries: "Africa , Asia, Europe, European Union, High income, International, Low income, Lower middle income, North America, Oceania, South America, Upper middle income, World"
    e. For GDP dataset, we keep only the following columns: "GDP, current prices (Billions of U.S. dollars, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021"
    f. For GDP dataset, we remove the rows with indices: "197 to 228". These rows are various regions of the world, like: Europe, Middle East, ASEAN-S, etc.
    g. For GDP dataset, we remove the rows with indices: "0, 229, 230". There are junk rows with no country names and data

2. Add any essential columns for the datasets:
    a. For Rio dataset, we add the column: "Total"

3. Rename the column name if required.
    a. For GDP dataset, we also rename the column: "GDP, current prices (Billions of U.S. dollars" to "Country".

4. Remove whitespaces around the column names and add underscore between the words of the column names.

5. Check for any missing values in the dataset.

6. Replace the integer missing value with 0.

a. In the "Covid Cases and Vaccinations" dataset, for the dates when the vaccination program was not started in a particular country, the values are marked as NaN. For our analysis, we are replacing the NaN records by zero (0).
b. The GDP dataset has some records with 'no data'. We replace them with 0 as these records do not have data.

7. Clean the country names by removing special characters, digits, and returning the Unicode normal form for the strings.

8. For the population data, we observe that the values are in the unit: thousands of persons. So, to get the exact figure, we multiply each of the records by 1000.

9. Country names may differ among the data collected from different sources. So, we identify the countries which may have different spellings in different datasets and modify the names to follow a common name. Our reference for this is the country names of the "Countries and Population" dataset.

After carrying out the above steps, the data is now stored in a MySQL database server in the tables and data format specified in the section 3.2 Description of Data for the datasets.

For the analysis of the data, we then query the required data from the relevant database tables and derive the desired insights.

Note: Jupyter Notebook ("*DMP_Data Processing and Storing.ipynb*"), detailing the steps for importing and cleaning the data and then storing it in a MySQL database is shared.

5. Data Analysis and Visualization

For our analysis we query our database and then plot the various charts to understand if there is any impact of covid-19 on the performance of the athletes participating in the Tokyo Olympics 2020 games.

First, we observe the performance of different countries who have participated in all the three summer Olympic games in 2012, 2016 and 2020. Then we identify the countries which have a drop in the number of medals earned.

Following charts are analysed:

1. country wise bar graph for total medals won in the three Olympic games.
2. country wise bar graph for gold medals won in the three Olympic games.
3. country wise bar graph for silver medals won in the three Olympic games.
4. country wise bar graph for bronze medals won in the three Olympic games.

Now, we attempt to identify if there is any correlation of the following factors in the performance of the countries for which we observe the change in the number of medals earned:

- Cumulative covid cases and new covid cases per population before the start of the games
- Cumulative deaths and new deaths per population before the start of the games
- Vaccinations per population carried out in the country before the start of the games
- Yearly value of GDP of the country

We plot a time series plot for the above metrices for some of countries for analysis.

We observe that in general, the countries whose GDP are higher than 900 billions of US dollars, were able to perform well in Tokyo Olympic-2020 compared to those countries who also had similar trends in above metrics but a lower GDP value.

Here we can postulate that, these countries' performance were not impacted due to covid-19 because of the athletes access to proper resources for training and preparation.

Finally, we are also providing an interactive option for an user to select a country, and then all the following details will be displayed for the selected country:

- Performance in all 3 Olympics: medal counts (total, gold, silver and bronze)
- Trend of cumulative covid and new cases per population till the start of Tokyo Olympic 2020.
- Trend of cumulative deaths and new deaths per population before the start of the Tokyo Olympic 2020.
- Vaccination per population trend till the start of Tokyo Olympic 2020
- Yearly value of GDP from the year 2012 to 2021.

The user can have a brief overview from the plotted charts of the above metrices for the selected country.

Note: Jupyter Notebook ("*DMP_Data Analysis and Visualisation.ipynb*"), detailing the steps for fetching, analysing and visualisation the data is shared.

6. Data Backup and Restore

Frequency of collection of our data sets is per year. Hence, we takes backup once new data is uploaded.

Using 'mysqldump' client utility program, we take logical backup of our database Olympic. The program produces set of SQL commands that can be executed to re-build the original database, along with the table defination and data. This type of backup is suitable for our smaller amounts of data and as backup stored in logical format is machine independent which makes it highly portable.

We are using "mysqldump" program to take backup and "mysql" for restore the database

If database crashes or its get corrupted, we can use this backup file to restore the database.

```
(base) tanuja@Tanujas-MacBook-Air bin % mysqldump -u root -p olympic > /tmp/backup_olympic_db.sql
Enter password:
(base) tanuja@Tanujas-MacBook-Air bin % cd /tmp
(base) tanuja@Tanujas-MacBook-Air /tmp % ls
backup.sql                      com.apple.launchd.FSmAAnIHIC    mysql.sock                      mysqlx.sock.lock
backup_olympic_db.sql           foo.err                         mysql.sock.lock                 powerlog
bitrock_installer.log           fseventsd-uuid                  mysqlx.sock
(base) tanuja@Tanujas-MacBook-Air /tmp % cd /usr/local/mysql/bin
(base) tanuja@Tanujas-MacBook-Air bin % mysql -u root -p olympic < /tmp/backup_olympic_db.sql
Enter password:
```

Figure 2: Procedure for Backup and Restore

Note: Backup file ("*backup_olympics_db.sql*") for our database is shared.

7. Data Access Control

As the data stored in the database for this project is publicly available knowledge and there is no sensitive data, so we provide read only access to all the database tables to an user.

Currently, the MySQL database is deployed in our local machine (personal laptop) and is not accessible remotely over any network.

However, if we deploy our database on a server (on premise or cloud server) which is connected over a network, then we will create a database user with read only privilege to access all the data tables stored in the 'olympic' database, using steps depicted in below picture.

```
mysql> CREATE USER 'data_access_user'@'%'
    -> IDENTIFIED BY 'p@ssw0rd';
Query OK, 0 rows affected (0.01 sec)

mysql> GRANT SELECT ON olympic.* TO 'data_access_user'@'%';
Query OK, 0 rows affected (0.01 sec)

mysql> SHOW GRANTS FOR 'data_access_user'@'%';
+-----------------------------------------------------+
| Grants for data_access_user@%                       |
+-----------------------------------------------------+
| GRANT USAGE ON *.* TO `data_access_user`@`%`        |
| GRANT SELECT ON `olympic`.* TO `data_access_user`@`%` |
+-----------------------------------------------------+
2 rows in set (0.01 sec)
```

Figure 3: Procedure to Create User to Access DB

8. References

1. https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html
2. https://dev.mysql.com/doc/refman/8.0/en/tutorial.html
3. https://dev.mysql.com/doc/refman/8.0/en/mysqldump.html
4. https://dev.mysql.com/doc/mysql-backup-excerpt/5.7/en/using-mysqldump.html
5. https://dev.mysql.com/doc/refman/8.0/en/connection-access.html