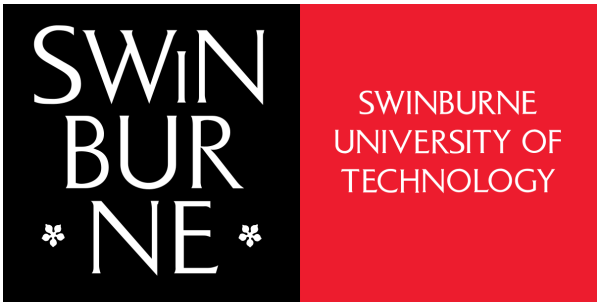


COMPREHENSIVE ANALYSIS OF THE 2020 TOKYO OLYMPICS: MEDAL DISTRIBUTION AND PARTICIPATION INSIGHTS

Distinction/High Distinction Report

COS80023 – BIG DATA

Report due: 27th October 2024

Pavan Kumar Muppala

103818552

ABSTRACT

This project examines the dataset from the 2020 Tokyo Olympic Games, including information on athletes, teams, disciplines, and coaches. I built the analysis using an Azure-based infrastructure, which involved Azure Data Factory for data ingestion, Azure Data Lake Gen 2 for storage, Azure Databricks for data transformation, and Azure Synapse Analytics for advanced querying and analytics. I also employed Tableau to visualize the data, providing clear and interactive ways to explore medal distribution by country, gender participation across various disciplines, and athlete performance in different sports. Using visualizations like stacked bar charts and map-based medal distribution, I made complex data patterns easier to understand. In this report, I describe the architecture, outline the data processing methods I used, and highlight the key insights revealed through Tableau's interactive visualizations. The project offers a comprehensive analysis of the Olympic dataset, providing valuable insights into athlete performance and participation trends.

TABLE OF CONTENTS

1 INTRODUCTION

This project analyzes the dataset from the 2020 Tokyo Olympics, focusing on uncovering key insights related to medal distribution, athlete participation, and gender trends across various disciplines. Using an Azure-based infrastructure that includes Azure Data Factory for ingestion, Azure Data Lake Gen 2 for storage, Azure Databricks for data transformation, and Azure Synapse Analytics for querying, I processed and analyzed the data efficiently. Tableau was employed to create interactive visualizations, making it easier to explore and interpret the data. The project aims to provide a clear understanding of the performance of countries and athletes during the 2020 Olympics, highlighting key trends in participation and medal distribution.

2 DATA DESCRIPTION

The dataset for the 2020 Tokyo Olympics includes comprehensive details about over 11,000 athletes participating in 47 disciplines and representing 743 teams. It contains key information such as the names, countries, disciplines, and gender of the athletes, as well as details about the coaches and teams involved. Additionally, the dataset includes entries by gender, providing insight into participation rates across various sports. This data serves as the foundation for analyzing trends in medal distribution, athlete performance, and gender participation. Data Source Link:

[2021 Olympics in Tokyo Dataset]

3 ARCHITECTURE OVERVIEW

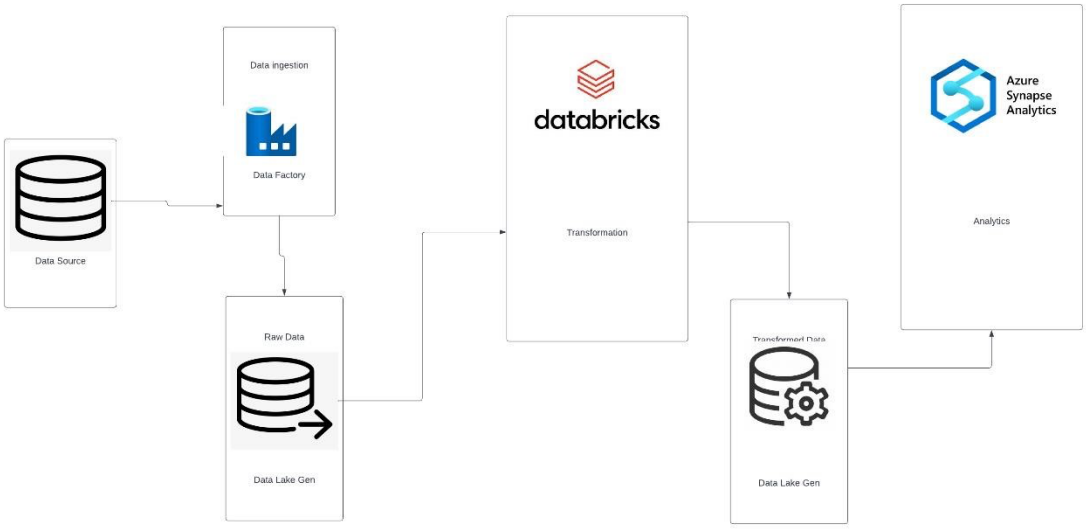


Figure 1: System Architecture

The architecture for the 2020 Tokyo Olympics data analysis project is built on an Azure-based pipeline to ensure smooth data processing, storage, and analysis. The flow starts with Azure Data Factory, which ingests the data from external sources. The raw data is then stored in Azure Data Lake Gen 2, providing a scalable storage solution.

The next step involves Azure Databricks, where data is transformed, cleaned, and aggregated to prepare it for analysis. Once processed, the transformed data is stored again in Azure Data Lake Gen 2 for long-term storage.

Finally, the processed data is loaded into Azure Synapse Analytics for advanced querying and analysis. This architecture supports efficient data handling and provides a seamless process from data ingestion to final analysis.

4 DATA PROCESSING AND TOOLS

The project involved multiple stages of data processing, utilizing a combination of Azure services and Tableau for visualization. Here is an overview of each step:

4.1 DATA INGESTION

In the Data Ingestion phase, I used Azure Data Factory to automate the process of loading various datasets related to the 2020 Tokyo Olympics into the Azure environment. The ingestion pipeline enabled the seamless extraction and transfer of raw data into Azure Data Lake Gen 2 for further transformation and analysis. This process ensured consistency, scalability, and efficiency in handling large datasets, minimizing manual intervention and streamlining the workflow for subsequent data processing stages.

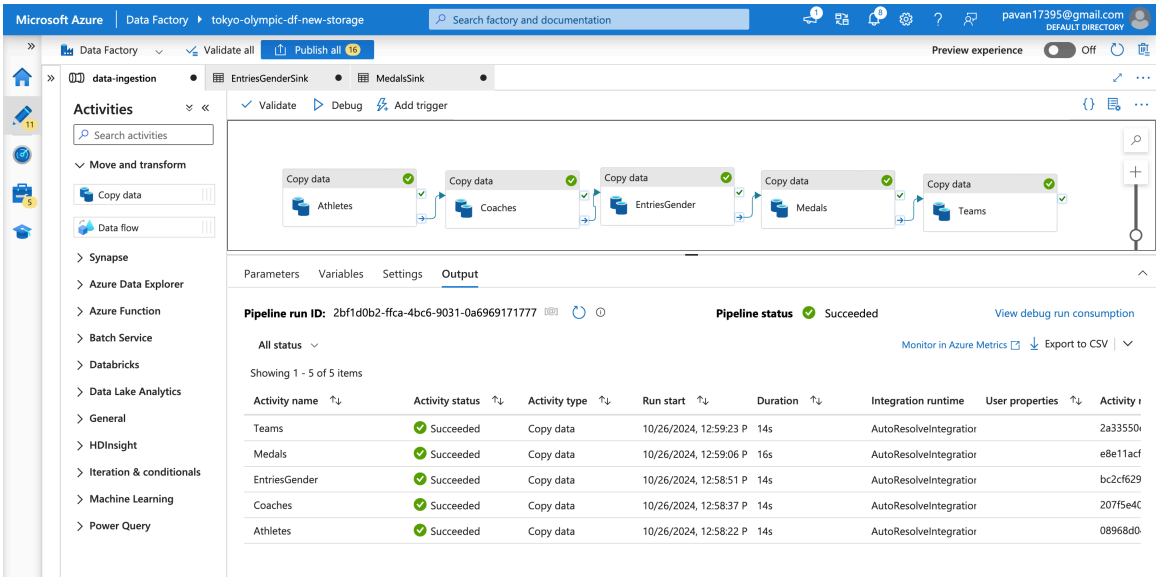


Figure 2: Loading data into Data Factory

4.2 DATA STORAGE

the Data Storage phase involved storing the ingested raw data in Azure Data Lake Gen 2. This storage platform ensured the scalability needed for large datasets and provided a secure space for both the raw and transformed data. Storing the data in this format allowed for easy access during the subsequent Data Transformation phase and maintained data integrity throughout the process. By using Azure Data Lake Gen 2, I also ensured that the data would be readily available for future queries and analysis, supporting long-term data management and retrieval. This storage step prepared the data for further cleaning and transformation in Azure Databricks.

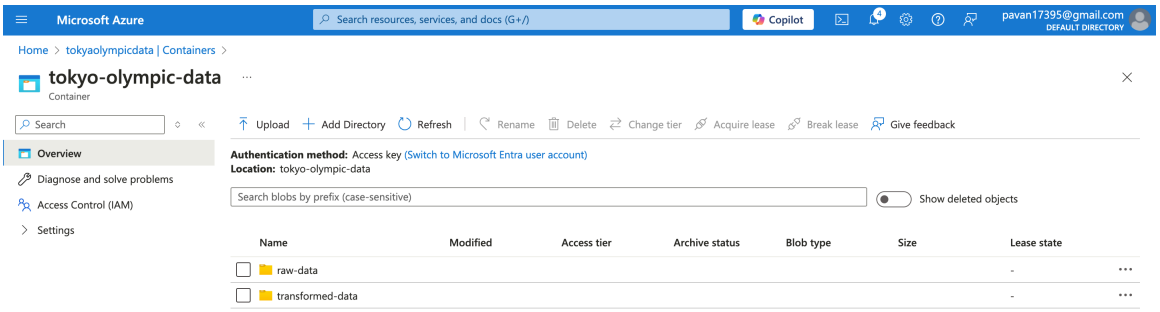


Figure 3: Data stored in Azure Data Lake Gen 2

4.3 DATA TRANSFORMATION

In the Data Transformation phase, I used Azure Databricks to handle a variety of data preparation tasks, such as data cleansing, transformation, and feature engineering. Azure Databricks offered a powerful environment for performing complex data operations efficiently. The process began with ingesting the 2020 Tokyo Olympics dataset, which comprised five CSV files: athletes, coaches, entriesgender, medals, and teams.

I utilized Spark's DataFrames API for loading and exploring the datasets. Each dataset was loaded into DataFrames using the spark.read.format("csv") method, with options like "header" and "inferSchema" set to automatically detect column names and data types. This allowed for easy exploration of the initial structure of the datasets.

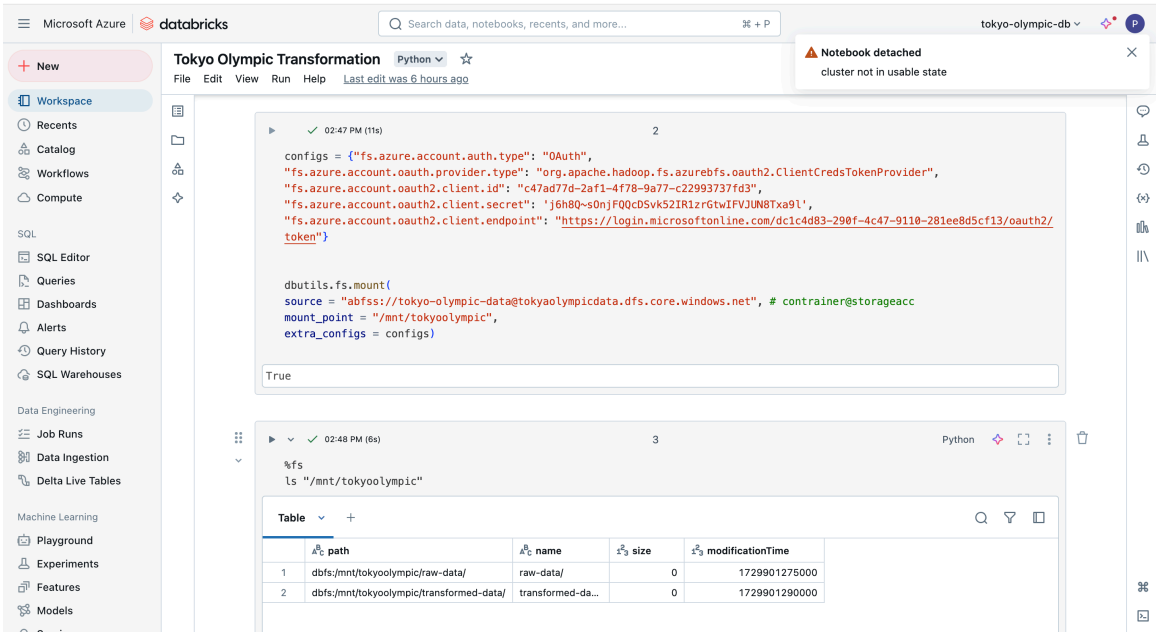


Figure 4: Azure Databricks

Additionally, I used the .withColumn() function to cast data types in the entriesgender DataFrame, ensuring consistency for calculations.

The transformation process included multiple steps aimed at extracting valuable insights from the data. For instance, I performed queries to identify the countries with the highest number of gold medals.

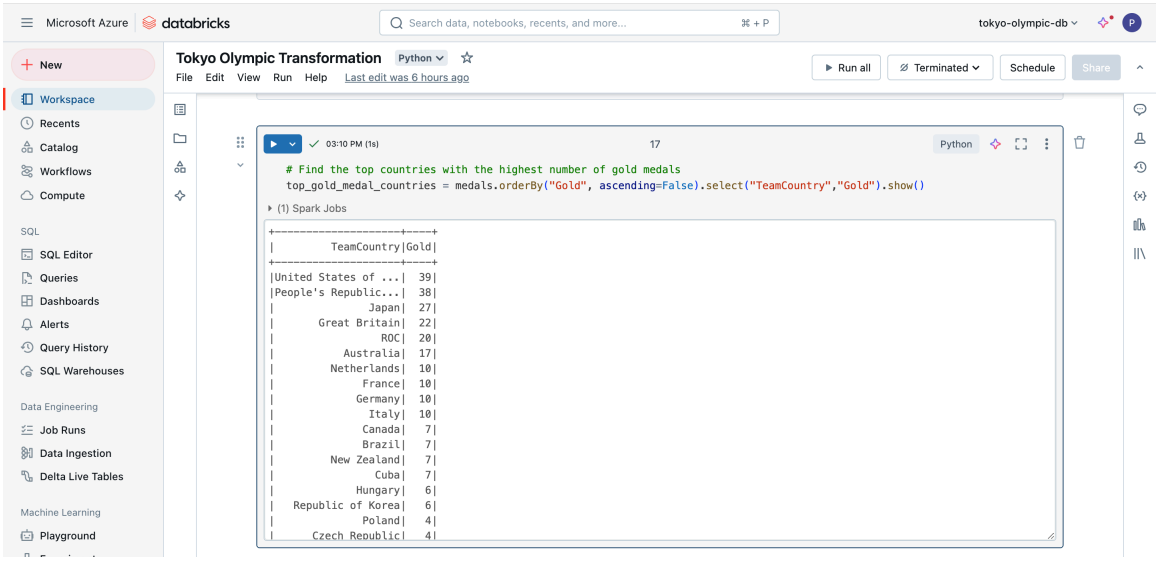


Figure 5: Countries with Highest Number of Gold medals

I also calculated the mean number of entries by gender for each discipline, improving the dataset's quality and making it ready for in-depth analysis.

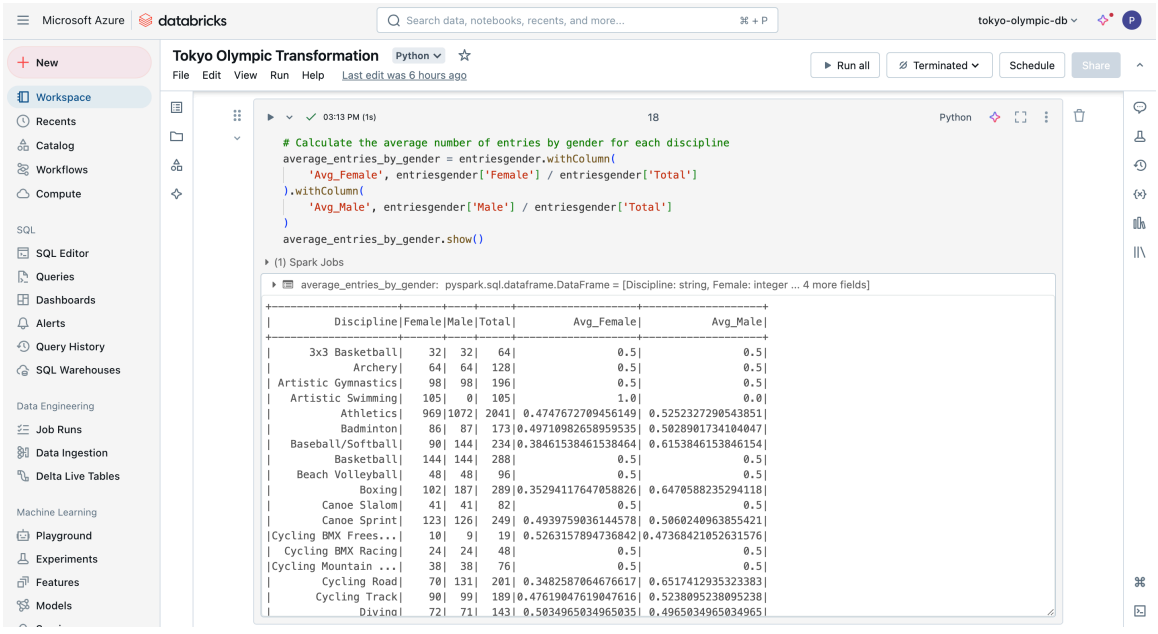


Figure 6: Average number of entries by gender for each discipline

After completing the transformations, I saved the processed data back into Azure Data Lake Gen 2, ensuring that the refined data was securely stored and well-organized. This structured approach in Azure Databricks streamlined the process of data transformation, making the dataset ready for further analysis and reporting.

4.4 DATA ANALYSIS

For the Data Analysis phase, I used Azure Synapse Analytics to perform advanced querying and reporting on the processed dataset. Synapse Analytics served as a powerful tool for extracting key insights, identifying

patterns, and generating detailed reports that summarized the performance and participation metrics from the 2020 Tokyo Olympics dataset.

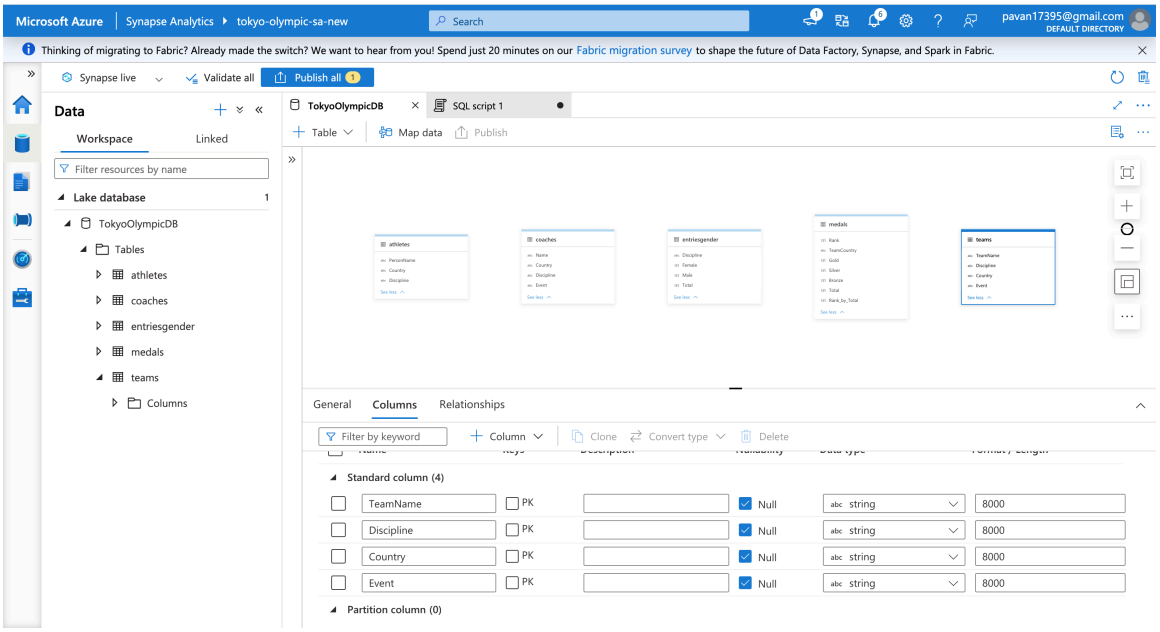


Figure 7: Created Azure Synapse Analytics and inserted all transformed data to perform advanced querying.

Using Synapse, I ran several queries to uncover valuable insights. For example:

- **Counting the number of athletes by country**: This query helped in understanding the distribution of athletes across participating nations.

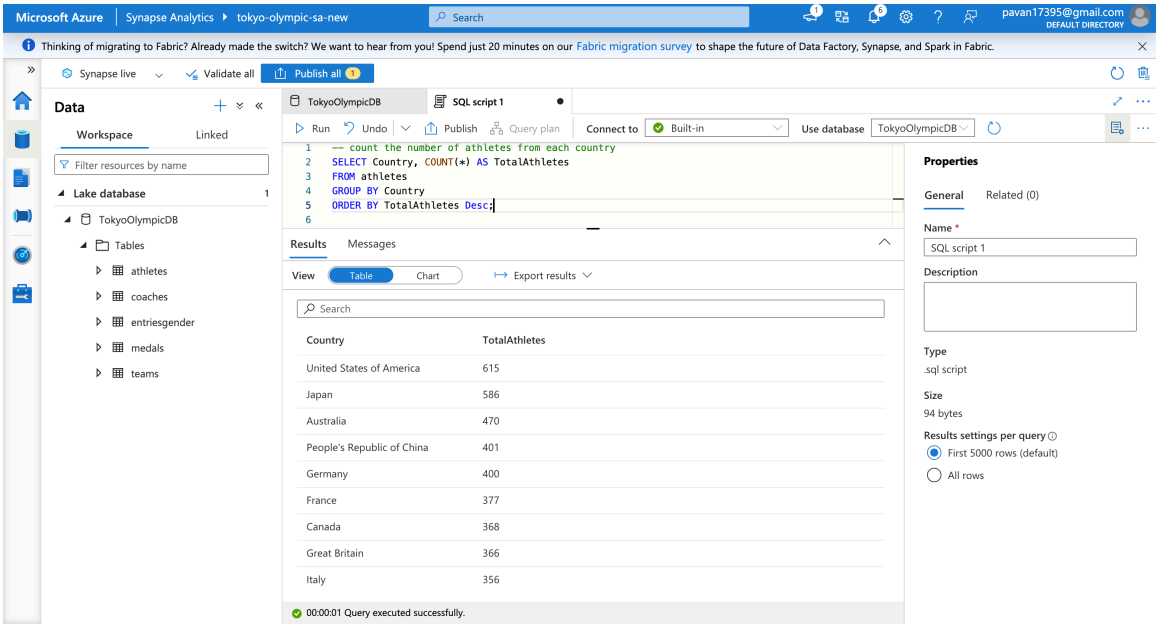


Figure 8: The Number of Athletes by country

- **Calculating the total medals won by each country**: By aggregating the medal counts (gold, silver, and bronze), this query highlighted the top-performing nations in the Olympics.

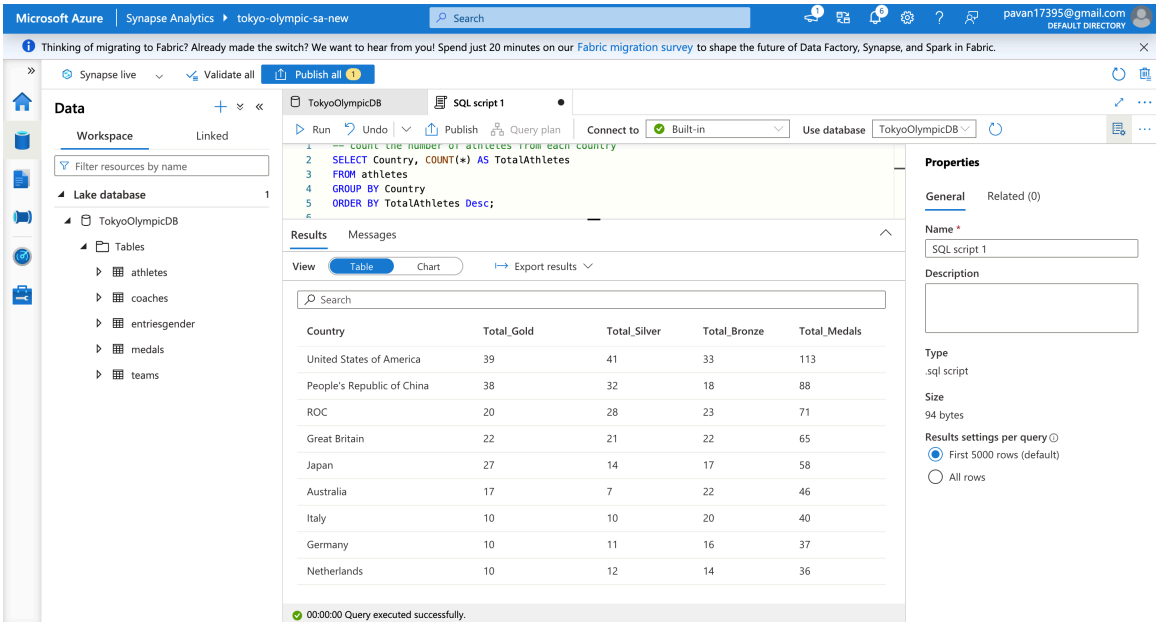


Figure 9: The total medals won by each country

- **Calculating the average number of entries by gender for each discipline**: This allowed for an analysis of gender participation across different sports, and the results were also visualized using line charts to illustrate trends.

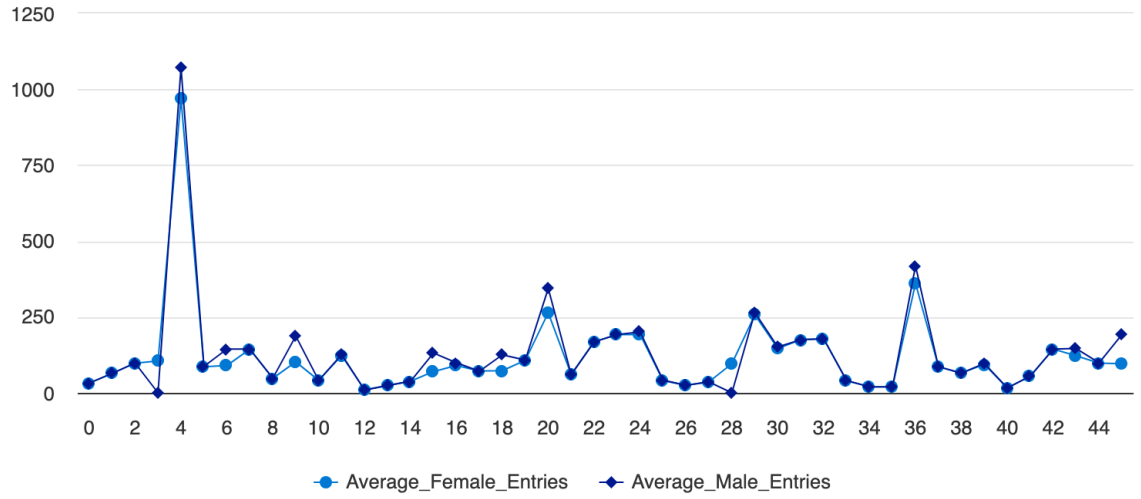
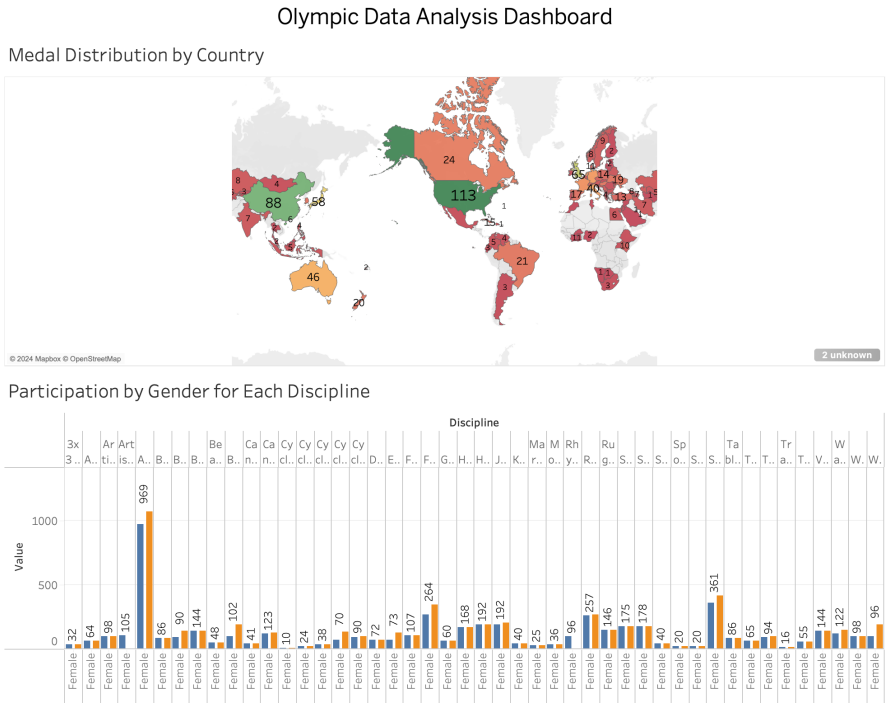


Figure 10: The average number of entries by gender for each discipline

This systematic approach, powered by Azure-based technologies, laid the groundwork for a thorough examination of the data. The combination of Exploratory Data Analysis (EDA) techniques, such as statistical analysis and visualizations, further uncovered essential insights into athlete demographics, team membership, and discipline performance. These findings were critical in understanding the broader trends within the 2020 Tokyo Olympics dataset and providing data-driven narratives.

4.4 DATA VISUALIZATION

In the Data Visualization phase, I used Tableau to create interactive charts and maps to highlight key insights from the 2020 Tokyo Olympics dataset. The visualizations include a medal distribution map that shows the total medal count for each country and a bar chart comparing male and female participation across various disciplines. These visuals provide a clear and intuitive way to explore patterns in medal performance and gender dynamics across different sports. The interactive nature of Tableau allows users to dive deeper into specific countries or disciplines for a more detailed analysis.



5 LIMITATION AND FURTHER RECOMMENDATIONS

Limitations:

- **Data Gaps and Inconsistencies**: Some datasets might contain missing values or inconsistencies, which can impact the accuracy and completeness of the analysis. This limitation requires additional data cleaning efforts.
- **Data Timeliness**: The data reflects the 2020 Tokyo Olympics and may not include updates or changes made after the event, limiting the scope for the most current analysis.

- **Absence of Real-Time Monitoring**: The project does not incorporate real-time data, which is crucial for obtaining live insights during ongoing events, such as future Olympic Games.
- **Limited Scope**: The analysis primarily focuses on medal distribution and participation by gender but does not cover other important factors, such as individual athlete performance, team dynamics, or historical comparisons.
- **Scalability Constraints**: Although Azure provides powerful tools, the scalability and compatibility of some services might pose challenges when dealing with larger datasets or more complex processing tasks.

Further Recommendations:

- **Data Enhancement**: Improve the dataset by addressing missing values and ensuring consistency through techniques like data validation and imputation. This will enhance the reliability of the analysis.
- **Real-Time Analysis**: Implement real-time data processing and analysis for future events, allowing for timely insights and more dynamic decision-making.
- **Ethical Data Management**: Ensure that all data is handled ethically, with a strong focus on privacy and compliance with data protection regulations, especially when dealing with sensitive personal information.
- **Adopt Advanced Tools**: Explore more advanced analytics tools and cloud platforms to optimize data processing workflows and enhance scalability.
- **Comprehensive Analysis**: Expand the analysis to include other factors, such as individual athlete performance metrics, team strategies, and historical trends across multiple Olympic events.
- **Collaborative Approach**: Engage with professionals from other disciplines, such as sports science and data privacy, to gain diverse perspectives and enhance the quality of the analysis.
- **Dashboard for Real-Time Monitoring**: Develop a live dashboard using Azure Synapse Analytics to provide real-time monitoring and reporting for future Olympic Games, enabling stakeholders to visualize and act on data as events unfold.

6 CONCLUSION

The analysis of the 2020 Tokyo Olympics dataset provided valuable insights into the performance and participation of athletes, the impact of team dynamics and coaching, and the role of gender in various sports.

Using a robust architecture based on Azure technologies, we processed, transformed, and visualized large datasets to uncover key patterns in medal distribution and athlete participation across disciplines.

Despite the limitations related to data quality, timeliness, and the absence of real-time monitoring, the analysis offered meaningful narratives about the event. The project also identified opportunities for future improvements, such as the inclusion of real-time dashboards, advanced analytics, and more comprehensive data.

This project highlights the significant role data plays in enhancing our understanding of large-scale sporting events like the Olympics. As data technologies continue to evolve, so too will our ability to extract deeper, more meaningful insights from global sports competitions, paving the way for richer analyses in the future.

REFERENCES

[1]Mishra, P. K., & Kumar, M. (2021). *Limitless Analytics with Azure Synapse*. Packt Publishing.

[2]Palacio, A. B. (2021). *Distributed data systems with Azure databricks : create, deploy, and manage enterprise data pipelines*. Packt Publishing.

[3]Rawat, S., & Narain, A. (2019). *Understanding Azure Data Factory : operationalizing big data and advanced analytics solutions*. Apress. https://doi.org/10.1007/978-1-4842-4122-6

[4]Stirrup, J., Nandeshwar, A., Ohmann, A., & Floyd, M. (2016). *Tableau: creating interactive data visualizations : illustrate your data in a more interactive way by implementing data visualization principles and creating visual stories using Tableau* (1st ed.). PACKT Publishing.