BUSINESS INTELLIGENCE PROJECT REPORT ON

"OLYMPICS DATASET ANALYSIS FROM YEAR 1896-2016 "

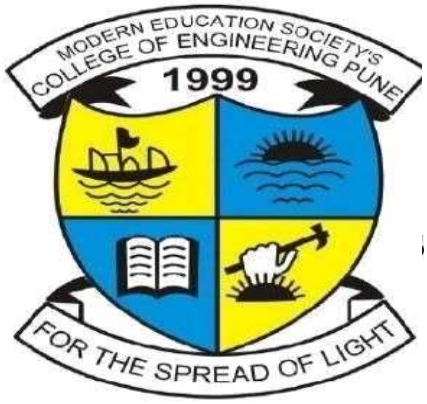
SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY,
PUNEIN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE
DEGREE OF

BACHELOR OF ENGINEERING (COMPUTER ENGINEERING)

BY

PARTH PIDADI	**S19111020**
PRATHAMESH NIKAM	**S19111028**

Under the guidance of
Prof. Jaya Mane



DEPARTMENT OF COMPUTER ENGINEERING
Modern Education Society's College of Engineering,
Pune, 19, BundGarden, V.K. Jog Path, Pune – 411001.

SAVITRIBAI PHULE PUNE UNIVERSITY 2022 – 23

CERTIFICATE

This is to certify that the project report entitles.

"OLYMPICS DATASET ANALYSIS FROM YEAR 1896-2016 "

Submitted by

PARTH PIDADI	**S19111020**
PRATHAMESH NIKAM	**S19111028**

is a bonafide work carried out by them under the supervision of **Prof. Jaya Mane** andit is submitted towards the partial fulfillment of the requirement of SAVITRIBAI PHULE PUNE UNIVERSITY, Pune for the award of the degree (Computer Engineering)

(Prof. Jaya Mane)	**(Prof. N.F. Shaikh)**
Guide	Head,
Department of Computer Engineering	Department of Computer Engineering

Place: Pune

Date:

ACKNOWLEDGEMENT

We would like to express our very great appreciation to **Prof. Jaya Mane** for her valuable and constructive suggestions during the planning and development of this research work. Her willingness to give her time so generously has been very much appreciated. Her patient guidance, enthusiastic encouragement, useful critiques, and her assistance in keeping our progression schedule throughout our semester will be acknowledged.

PARTH PIDADI S19111020
PRATHAMESH NIKAM S19111028

ABSTRACT

Data is one of the most essential commodities for any organization in the 21st century. Harnessing data and utilizing it to create effective marketing strategies and making better decisions is extremely essential for organizations. For a sporting event as big as Olympics, it is necessary to organize and analyze the large volumes of data generated to make sense of existing performance and identify growth potential. This project is been performed to understand how different players and countries performed in various sporting events of the Olympics game.

The main motive of doing this project is to graphically and analytically describe how Olympics events had evolved over the time, using Power BI and Jupyter Notebook respectively. This project analyses many factors of Olympics such as diversity of participation, performance of athletes and different sports and events.

Olympics data analysis allows us to evaluate and compare the performance of athletes, teams and countries. By analyzing factors such as timings, scores, distances, and other performance metrics, we can assess the athletes' strength, weaknesses and overall progress.

An exploratory data analysis has been performed onthe dataset to explore the various factors such as number of male and female participants, total number of medals won i.e. Gold, Silver or Bronze by individuals or by countries etc. Additionally, a dashboard highlighting information about event has been created in Power BI and provides an overview of the overall performance of various individuals and countries in various sports.

TABLE OF CONTENTS

1. INTRODUCTION

The Olympic Games, the ultimate stage of athletic competition, capture the world's attention with their splendor, sportsmanship, and extraordinary displays of human achievement. Athletes from all corners of the globe gather every four years, driven by the desire for glory, record-breaking performances, and the opportunity to etch their names into the annals of sporting history. While the physical prowess showcased during the Games is awe-inspiring, the era of data-driven insights has revealed that analyzing Olympic athlete events goes beyond sheer athleticism. By delving into the vast repositories of data generated during the Olympics, we can uncover the intricate details of athlete performances, discern trends, identify patterns, and extract invaluable insights that shape coaching strategies, talent identification, rule evaluation, fan engagement, and the future of Olympic sports.

The analysis of Olympic athlete events provides a multi-dimensional approach to understanding the true essence of sporting excellence. It transcends mere medal counts and podium finishes, delving into the very fabric of athletic achievement. Athletes compete across a myriad of disciplines, each presenting its unique challenges, skill sets, and strategic intricacies. From the lightning speed of the track to the precision of gymnastics, the power of weightlifting to the grace of figure skating, the Olympics serve as a culmination of years of training, determination, and unwavering commitment.

In the age of technological advancements, data has permeated every facet of our lives, transforming industries and revolutionizing decision-making processes. The world of sports is no exception. Within the vast sea of data generated during the Olympics lies a treasure trove of insights waiting to be unearthed. Every recorded second, every leap, every throw, and every stroke captures a moment frozen in time, ready for analysis and interpretation. Through rigorous examination, this data can yield valuable information that shapes the future of athletic performance.

The analysis of Olympic athlete events yields numerous benefits that extend beyond the realm of sport. Coaches and trainers can delve into the intricacies of an athlete's performance, meticulously dissecting each component to identify areas for improvement, strengths to be maximized, and weaknesses to be addressed. This data-driven approach

enables evidence-based coaching strategies, informed training programs, and personalized guidance tailored to the unique needs of each athlete.

Moreover, the analysis of Olympic athlete events plays a pivotal role in talent identification and development. In a world brimming with potential sporting prodigies, pinpointing exceptional talents can be a daunting task. By studying the data, patterns, and emerging trends from the Olympics, talent scouts, national federations, and sports academies can identify promising athletes and nurture their potential. This ensures that talent is not only discovered but also provided with the necessary resources, guidance, and support to flourish on the global stage.

Beyond performance evaluation and talent identification, the analysis of Olympic athlete events has far-reaching implications for rule evaluation and equipment innovation. The Olympics often witness evolutions in sports rules, as well as advancements in equipment design and technology. By subjecting the data to thorough scrutiny, officials and governing bodies can evaluate the impact of rule changes, assess fairness, and ensure the integrity of the sport. Similarly, analyzing the performance of athletes with different equipment variations aids in determining the effectiveness and impact of such advancements on performance.

In addition to practical implications, the analysis of Olympic athlete events holds significant value for fan engagement and storytelling. The Games, with their rich history, iconic moments, and unparalleled emotions, captivate audiences worldwide. Data analysis adds another layer to the narrative, allowing fans to delve into the statistics, historical records, and insightful visualizations. These data-driven insights provide a deeper understanding of the context, history, and significance of different events and performances, creating a more immersive and engaging experience for viewers.

The analysis for this study has been done using SQL, R, Python, and Power BI on the dataset provided by Olympic organization on Kaggle (athletes_events and noc_regions dataset). The modeling, as well as the exploratory data analysis for the research, have been performed in R and Python, aggregation and querying will beperformed using SQL and the final dashboard has been created using Power BI.

1.1 TOOLS AND TECHNOLOGIES APPLIED

The analysis for this study has been performed using some main tools: R, Python, and Power BI. The models and Exploratory Data Analysis have been executed using development tools like R Studio and Jupyter Notebook

Several packages have been used to perform the initial and final outcome EDA for the analysis. For the initial EDA, a combination of R and Python libraries like inspectdf, ggplot2, plotly, caret, matplotlib, seaborn, etc have been implemented. Packages like numpy, pandas, tidyverse, etc. have been used for data wrangling and manipulation. For the models that have been created, several packages like 'scikit-learn', 'xgboost', etc have been applied.

2. PROBLEM STATEMENT

The Olympic Games stand as the pinnacle of sporting achievement, showcasing the world's most exceptional athletes competing on a global stage. Behind these awe-inspiring performances lies a wealth of data waiting to be explored and analyzed. However, the full potential of Olympics datasets remains untapped, and there is a pressing need to harness this information to enhance athlete performance and unlock new avenues for success.

Despite the wealth of data generated during the Olympics, there are challenges in effectively utilizing and interpreting this vast repository of information. Coaches, trainers, and sports organizations struggle to extract meaningful insights from the data, hindering their ability to develop targeted training programs and identify areas for improvement. Without a comprehensive understanding of athlete performances, it becomes difficult to optimize training regimens, tailor strategies, and unlock the full potential of each individual.

Additionally, the sheer volume and complexity of the data present challenges in data management, integration, and analysis. Traditional methods of data analysis are often insufficient to handle the scale and intricacies of Olympics datasets. Extracting meaningful patterns, identifying correlations, and making accurate predictions requires advanced analytics techniques and specialized tools.

Furthermore, the lack of standardized data formats across different Olympic events and sports disciplines poses a significant obstacle to comprehensive analysis. Datasets from various competitions may differ in structure, variables, and data collection methodologies, making it challenging to compare and combine information. This fragmentation hampers efforts to derive overarching insights and limits the ability to identify universal patterns of success.

To address these challenges, there is a need to develop robust methodologies and frameworks for Olympics dataset analysis. This involves leveraging advanced analytics techniques, machine learning algorithms, and data visualization tools to uncover hidden patterns, trends, and relationships within the data. By integrating diverse datasets and establishing standardized formats, researchers and practitioners can gain a holistic view of athlete performances, enabling evidence-based decision-making and targeted interventions.

Moreover, there is a need to bridge the gap between data analysis and practical application in the realm of sports. The insights derived from Olympics dataset analysis must be translated into actionable recommendations for coaches, trainers, and athletes. By establishing a feedback loop between data analysis and performance enhancement, sports organizations can drive continuous improvement, optimize training approaches, and unlock the full potential of athletes.

In summary, the problem statement for Olympics dataset analysis revolves around the need to harness the power of data to enhance athlete performance and maximize their potential. By overcoming challenges related to data integration, analysis methodologies, and practical application, researchers and practitioners can unlock valuable insights, develop evidence-based strategies, and drive performance improvements in the world of Olympic sports.

3. METHODOLOGY

The project comprises of analyzing the Olympics dataset whose methodology is as follows:

1. Data Preparation: The first step is to prepare the dataset for analysis. This includes importing the dataset into Jupyter Notebook, checking for missing values, and removing duplicates if any. The dataset can be read into Jupyter Notebook using the Pandas library.

2. Data Exploration: Once the dataset is loaded, the next step is to explore the data. This involves examining the different features of the dataset such as the number of athletes, countries, and sports. Descriptive statistics can be used to summarize the data and identify any patterns or trends. This can be achieved using Pandas' describe () function.

3. Data Visualization: After exploring the data, the next step is to visualize it using charts and graphs. This can be done using Python visualization libraries like Matplotlib and Seaborn. Visualizations can help identify trends and patterns in the data that may not be immediately apparent from just exploring the dataset.

4. Data Analysis: Once the data is prepared, explored, and visualized, the next step is to analyze it. This can include examining the performance of specific countries, analyzing the distribution of medals across sports and disciplines, or identifying factors that contribute to success in a particular sport. The analysis can be performed using Python's scientific computing libraries like NumPy and SciPy.

5. Interpretation and Conclusion: The final step is to interpret the analysis and draw conclusions from it. This involves synthesizing the findings from the analysis and identifying the implications of the results. The conclusions drawn from the analysis can help inform future decision-making processes and strategies for the Olympic Games.

4. ABOUT THE DATASET

The dataset for this study has been acquired from a past Kaggle. This can be found here: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results. This data set is a result of an incredible amount of research by a group of Olympics history enthusiast, which contains information about individual. It contains information about individual athlete competing to an individual Olympic events.

The 'athlete_events.csv' data file has several special attributes to specify each athlete. For example: ID – A unique number for each athlete, the ID column can be used to uniquely identify athletes, since some athletes have same name. The dataset also contains information about the events, participated teams, medals etc.

The main goal of this study is to extract valuable insights and patterns from the vast amount of data generated during the Olympic Games. By analyzing various aspects such as athlete performance, event results, historical trends, and audience engagement, data analysis aims to provide a deeper understanding of the games. It helps in identifying factors that contribute to success, evaluating the impact of various variables, optimizing training strategies, enhancing decision-making processes, improving event management, and uncovering opportunities for innovation and improvement. Ultimately, the goal is to leverage data to enhance the overall Olympic experience for athletes, organizers, and spectators alike.

In total, there are 2,71,116 rows in the dataset and 15 columns in the athlete_events dataset. (Figure 1)

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
676	Gaston Achille	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Individual	NA
757	Adam	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Individual	NA
4073	Andreac	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Individual	NA
4204	Emmanuel Andrieu	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Masters, Individual	NA
5684	Ass	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Masters, Individual	NA
5938	Aufort	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Masters, Individual	NA
5938	Aufort	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Sabre, Masters, Individual	NA
8973	Bazin	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Individual	NA
9728	G. Bllot	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Individual	NA
10811	Charles Bersin	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA
10811	Charles Bersin	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Sabre, Masters, Individual	NA
11178	Raoul Bzy-Bideau	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Masters, Individual	NA
11329	Raoul Bideau	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Individual	NA
13551	Bormel	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA
13551	Bormel	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Masters, Individual	NA
13903	mile Bouard	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA
14029	Marcel Boulanger	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA
14035	Jean A. Boulge	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA
14712	J. Brassard	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA
14712	J. Brassard	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Masters, Individual	NA
14741	Brau	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA
17274	Albert Simon Cahen	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Individual	NA
17274	Albert Simon Cahen	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's epee, Individual	NA
17521	Calvet	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Individual	NA
17644	Camier	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Sabre, Masters, Individual	NA
17929	Cannesson	M	NA	NA	NA	France	FRA	1900 Summer	1900	Summer	Paris	Fencing	Fencing Men's Foil, Masters, Individual	NA

Figure 1. A summary of the athlete_events.csv dataset

There is another dataset called ‘noc_regions.csv’ that contains some more detailed information about the code of name of country and its code. It has 230 rows and 2 columns.

NOC	Country
AFG	Afghanistan
AHO	Curacao
ALB	Albania
ALG	Algeria
AND	Andorra
ANG	Angola
ANT	Antigua
ANZ	Australia
ARG	Argentina
ARM	Armenia
ARU	Aruba
ASA	American Samoa
AUS	Australia
AUT	Austria
AZE	Azerbaijan
BAH	Bahamas
BAN	Bangladesh
BAR	Barbados

The datasets have been cleaned and analyzed using Jupyter Notebook andthe final Power BI dashboard has been created based on these values.

4.1 EXPLORATORY DATA ANALYSIS

It is crucial to have an in-depth understanding of the dataset that is used in this anal- lysis to derive accurate insights. Several times there are underlying patterns or trends in the data that would not be identified as easily, hence the need for an extensive exploratory data analysis. This thorough exam- inaction is necessary to understand the underlying structure of the dataset and to draw conclusions orinsight about the validity of our analysis.

The study is going to begin with a brief analysis of the available dataset to get a sense of the main characteristics and components that are relevant to the research. An exploratory data analysis is crucial to this study considering the numerous attributes that are a part of the dataset that will be essential when trying to draw insights and making predictions. As part of the exploratory data analysis, several visualizations have been created that will help us understand what it is that we are trying to achieve and to keep in mind the various attributes that we can use to improve results.

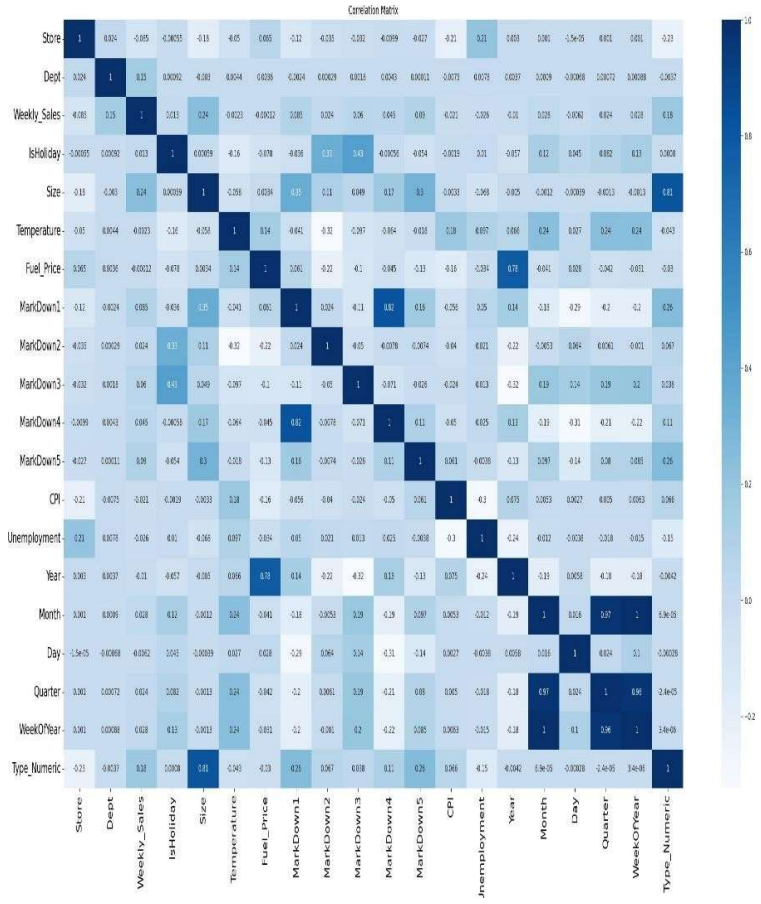
The EDA is like a primary investigation and tries to look at the relationships and nature of the different columns available to us. As part of this, the 'inspectdf' pack- age (Ellis, 2019) and the 'glimpse' package (Sullivan, 2019) have been used and imple- mented in R that will answer questions related to the number and nature of columns and rows in the dataset, missing values, distribution of numeric and categorical vari- ables, correlation coefficients, etc.

Several other packages like 'ggplot2', 'matplotlib', 'seaborn', and 'plotly' have also been used in this study to create visualizations that provide information about the Medal count data considering various aspects such as medals won by male from and to a particular period, total medal won by an individual, etc and plot them in tree map, card, matrix. These visualizations areaccompanied by brief descriptions that will discuss the findings and scope.

4.2 CORELATION MATRIX

A correlation matrix describes the correlation between the various variables of a dataset. Each variable in the table is correlated to each of the other variables in the table and helps in understanding which variables are more closely related to each other .

With the numerous variables available through this dataset, it became imperative to study correlations between some of them. By default, this matrix also calculates correlation through Pearson's Correlation Coefficient that calculates the linear relationship between two variables, within a range of -1 to $+1$. The closer the correlation to $|1|$, the higher the linear relationship between the variables and vice versa.



5. DATA CLEANING AND PREPROSESSING

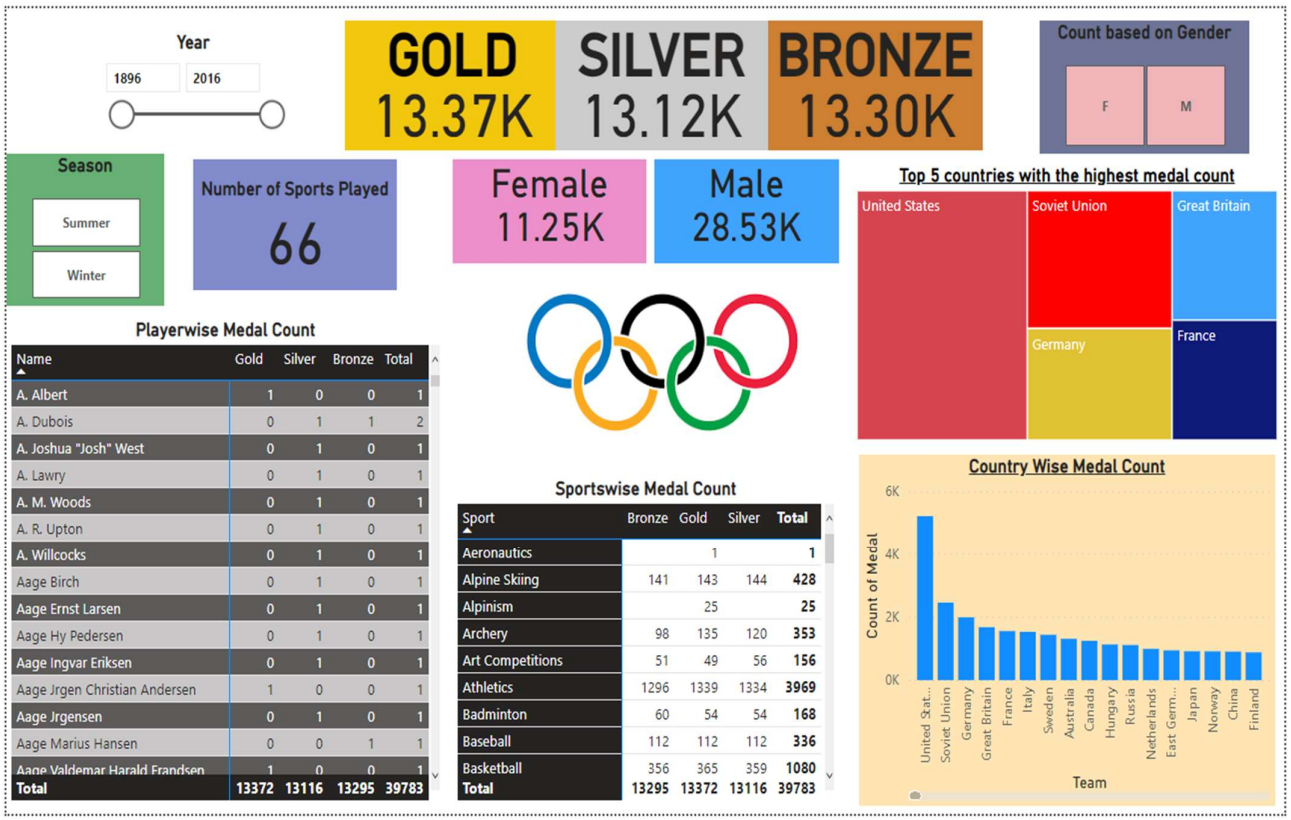
Data cleaning and preprocessing are essential steps in analyzing any dataset, including the Olympics athlete_events.csv dataset. the below steps were performed on the dataset
Handling Missing Values: Identify missing values in the dataset and decide on an appropriate strategy for handling them. This can involve imputing missing values based on statistical measures like mean or median, or removing rows or columns with excessive missing values if they are not significant.

a.] Standardizing Formats: Ensure consistent formatting of data across columns. This includes standardizing date formats, capitalization, and units of measurement.
b.] Handling Outliers: Identify and handle outliers, which are extreme values that may significantly affect the analysis. Outliers can be treated by either removing them if they are erroneous or by transforming them using techniques like Winsorization.
c.] Encoding Categorical Variables: Convert categorical variables into numerical representations that can be used for analysis. This can involve techniques such as one-hot encoding, label encoding, or ordinal encoding.
d.] Feature Scaling: Normalize or scale numerical features to bring them to a similar range. Common scaling techniques include min-max scaling or standardization.
e.] Dealing with Inconsistent Data: Examine the dataset for inconsistent or conflicting information and resolve any discrepancies. This could involve cross-checking data from different sources or validating against predefined criteria.
f.] Feature Selection: Identify relevant features for analysis based on domain knowledge or statistical techniques. Removing irrelevant or redundant features can improve the efficiency and g]. interpretability of the analysis.
h.] Handling Skewed Data: If the dataset exhibits a skewed distribution, applying transformations like logarithmic or power transformations can help normalize the data and improve the analysis.
I]. Data Integration: If additional datasets or external data sources are available, integrate them with the Olympics dataset to enhance the analysis with more comprehensive information.
By performing these data cleaning and preprocessing steps, the Olympics athlete_events.csv dataset was prepared for more accurate and meaningful analysis.

6. BUILDING BI DASHBOARD

As an end product, this Power BI dashboard is going to serve as the final product of this research. The dashboard contains detailed information about the original data related to the Olympics events conducted since 1896 to 2016.

The dashboard can be found in the final submitted folder. If a user does not have accessto Power BI, a PDF export of the entire dashboard is included along with the .pbix file that contains all of the created visualizations and reports in the dashboard.View of the dashboard created is included below:



7. CONCLUSION.

Concluding an analysis of the Olympics dataset from 1896 to 2016 requires a detailed examination of the dataset and the specific analyses conducted. Without specific analysis objectives and findings, a generalized conclusion can be made regarding the trends and patterns observed in the dataset over this period. However, it is important to note that the following conclusion is a broad summary and may not encompass the entirety of insights that can be derived from the dataset.

The analysis of the Olympics dataset from 1896 to 2016 reveals several notable trends and patterns. It showcases the growth and evolution of the Olympic Games, with an increasing number of participating countries and athletes over time. The dataset highlights the dominance of certain nations in specific sports or disciplines, with traditional powerhouses like the United States, Russia/Soviet Union, and Germany consistently performing well.

Additionally, the dataset provides insights into the changing demographics of Olympic participants, with a significant increase in female participation and the inclusion of new sports over the years. It also showcases the impact of geopolitical events, such as World Wars and political boycotts, on the participation and outcomes of the Olympic Games.

By analyzing the dataset, researchers can gain a deeper understanding of athletic performance trends, medal distributions, and the success of different countries across various Olympic Games. It allows for the identification of sports or disciplines where specific nations excel and provides insights into the factors contributing to their success.

Overall, the analysis of the Olympics dataset from 1896 to 2016 offers valuable insights into the historical development, achievements, and trends of the Olympic Games. These insights can inform future decision-making processes, including strategies for athlete development, event planning, and promoting inclusivity and diversity in the Olympic movement.

8. REFERENCES

1. R. Nand, A. Chand and E. Reddy, "Comparative Analysis of Classification Techniques on Olympic Games Datasets," 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2022, pp. 1-6, doi: 10.1109/CSDE56538.2022.10089334.
   https://ieeexplore.ieee.org/abstract/document/10089334
2. V. Asha, S. P. Sreeja, B. Saju, N. C S, P. G. N and A. Prasad, "Performance Analysis of Olympic Games using Data Analytics," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1436-1443, doi:10.1109/ICEARS56392.2023.10084943.
   https://ieeexplore.ieee.org/abstract/document/10084943
3. S. Xinyi and X. Chenglong, "Visual analysis of the distribution characteristics and influencing factors of Olympic medals," 2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2022, pp. 1633-1637, doi: 10.1109/IMCEC55388.2022.10019867.https://ieeexplore.ieee.org/abstract/document/10019867
4. Y. Wang et al., "DataShot: Automatic Generation of Fact Sheets from Tabular Data," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 895-905, Jan. 2020, doi: 10.1109/TVCG.2019.2934398.
   https://ieeexplore.ieee.org/abstract/document/8805442