**Olympic Data Analysis Report**

**Student Name:** Muhammad Naveed Anjum  
**Student Number:** 23026267  
**GitHub Repository:**

**Introduction**

This report analyzes Olympic data sourced from Kaggle, focusing on exploring and explaining relationships within the dataset. The dataset comprises records of athletes' performances across different Olympic games, including age, gender, team, event, and medal outcomes.

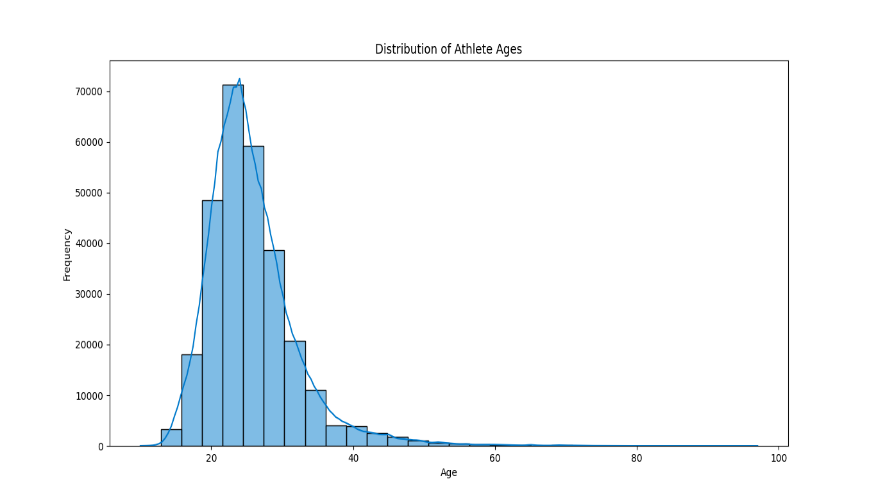
**Data Cleaning and Preparation**

The initial step involved merging athlete performance data with region information to enrich our dataset. Key data cleaning actions included:

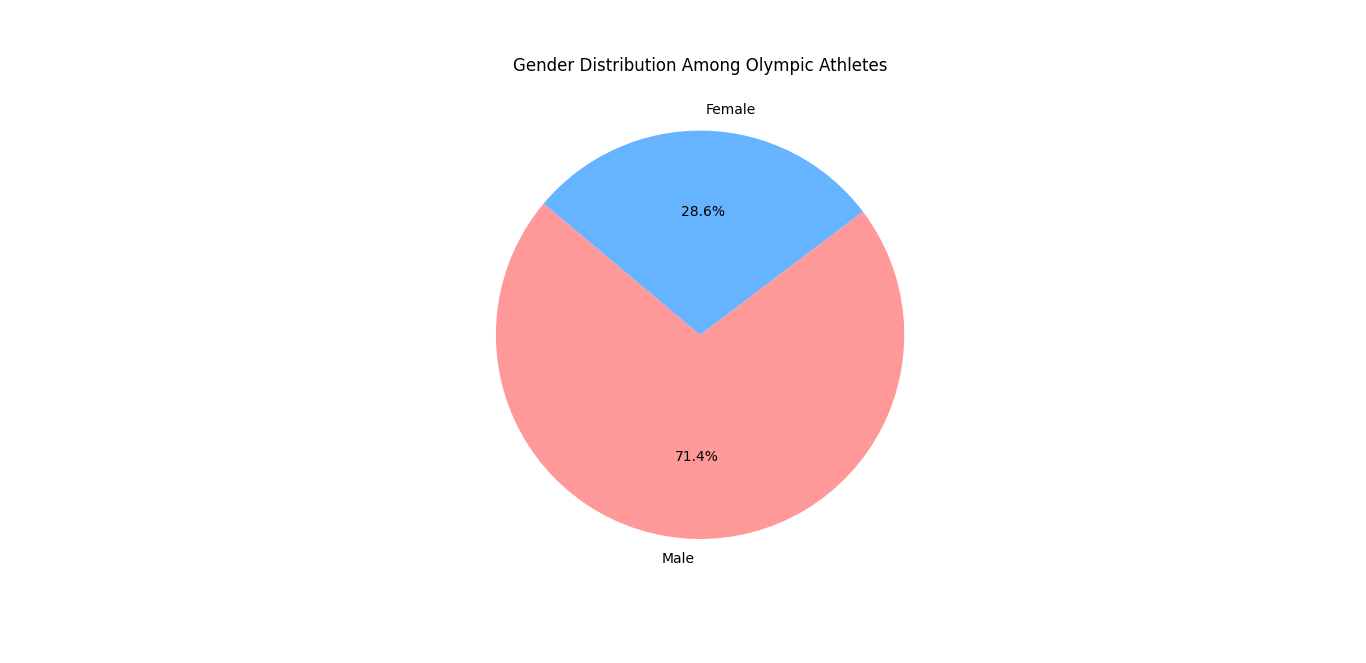
* **Age Imputation**: Missing ages were imputed using the median age of athletes within the same sport and event categories, reducing missing values from 9,474 to 142.
* **Medal Indicator**: A binary indicator was created to distinguish between medalists and non-medalists, facilitating analysis of medal-winning performances.
* **Handling Missing Values**: Missing region data for 23 entries was noted but not imputed due to complexity; this aspect could be addressed in future analyses.

**Exploratory Data Analysis and Visualization**

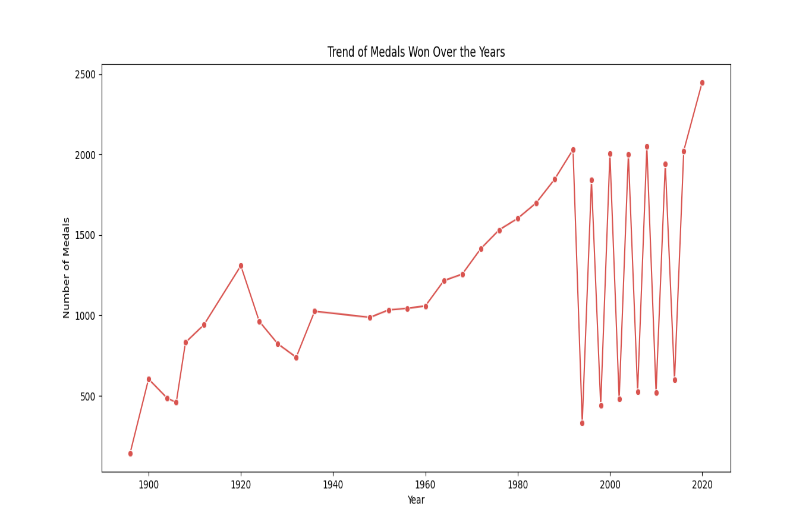
**Distribution of Athlete Ages**

The histogram titled "Distribution of Athlete Ages" depicts a right-skewed distribution, indicating a higher concentration of younger athletes with the mode being under 30 years old and a gradual decline in frequency as age increases. Despite the majority being younger, the long tail extending towards older ages suggests the presence of older athletes, potentially as outliers. The peak frequency surpasses 60,000, reflecting a substantial sample size and implying that the distribution is likely representative of the athlete population. The uniform bin width maintains the integrity of the frequency representation, highlighting that while there is a broad range of ages, the population is predominantly young with diminishing representation as age progresses.****

The pie chart titled "Gender Distribution Among Olympic Athletes" illustrates a significant gender imbalance among athletes, with males representing a majority at 71.4% and females at 28.6%. This visual representation, using a stark contrast in segment size and color coding, effectively communicates the disparity between male and female participation rates in the Olympics. The chart simplifies the data into a format that is immediately understandable, highlighting the need for deeper analysis into the underlying causes of this imbalance and its implications for gender equity in sports.

****

**The trend of Medals Won Over the Years**

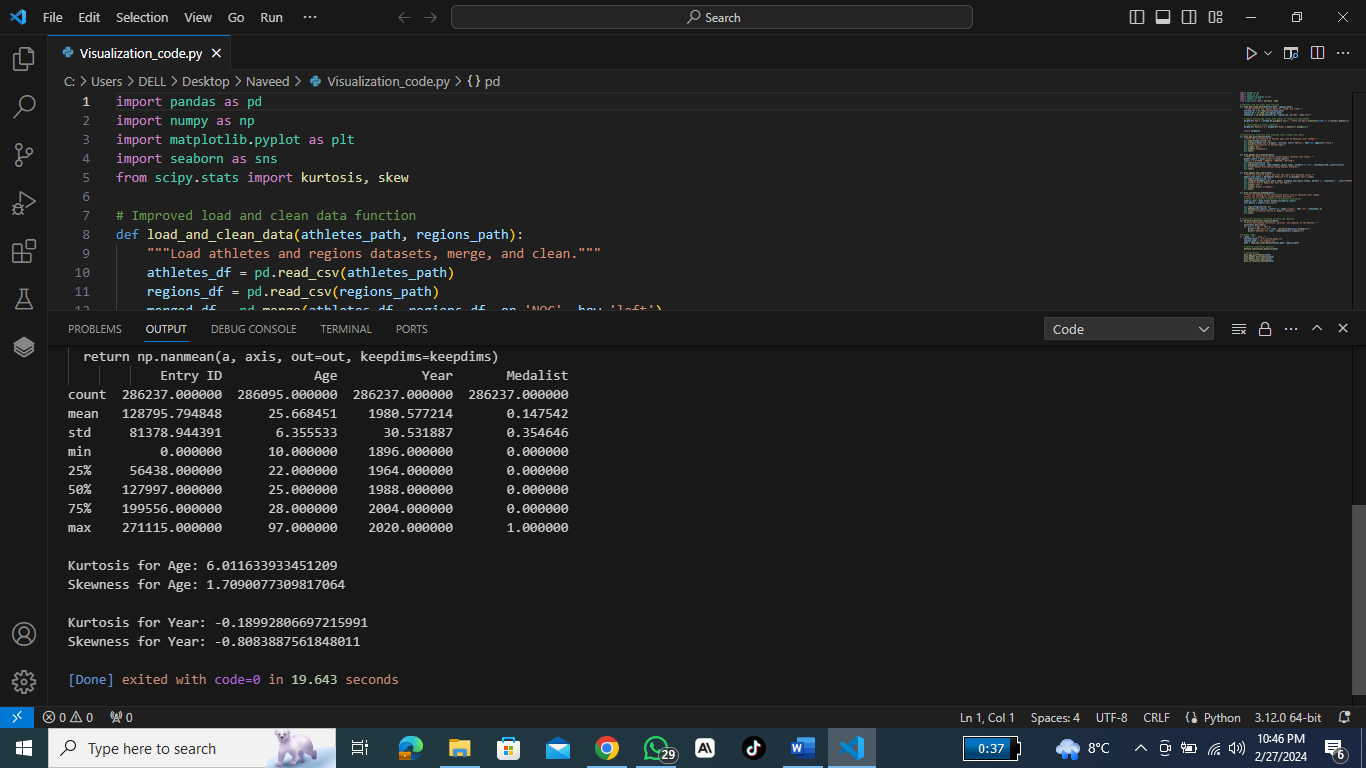
The graph "Trend of Medals Won Over the Years" illustrates the evolution of medals awarded in the Olympic Games from 1900 onwards. It exhibits a general increase in the number of medals over time, particularly after the 1960s. Intermittent sharp decreases likely indicate years when the Olympics were not held, presumably due to major global events like wars. These dips are immediately followed by recoveries, suggesting temporary disruptions rather than a true decline in medals awarded. The overall trend suggests a broadening of the Games, potentially due to the inclusion of new sports, more events, and an increase in participating nations.****

**Correlation Matrix of Numeric Features**

****The heatmap illustrates a correlation matrix with weak correlations among the numeric features: Entry ID, Age, Year, and Medalist. Entry ID is not significantly correlated with any other feature. Age and Year show a very slight negative correlation, as does Year with Medalist status, both of which are so weak as to be negligible. The chance of being a Medalist shows almost no correlation with Age or Entry ID, suggesting independence from these variables. Overall, no strong correlations are evident from this matrix.

**Statistical Summary**

The analysis utilized pandas' .**describe()** function to summarize the dataset's central tendencies, dispersion, and shape of the distribution. The **corr()** function provided a correlation matrix, highlighting relationships between numeric variables, such as age and year of competition, which were further visualized in the heatmap.



**Critical Assessment**

The exploratory data analysis revealed insightful trends and distributions within the Olympic dataset. Age distribution indicated a prime range for Olympic athletes, with gender distribution reflecting historical participation trends. The analysis of medals won over the years suggests an increase in competitiveness and inclusivity in the Olympic Games.

However, the missing data, particularly in the age and region fields, posed challenges. The imputation strategy for age was effective but imperfect, indicating the need for more sophisticated techniques or more complete datasets in future studies. The lack of region data for some athletes also suggests a potential area for data enrichment.

**Conclusion**

This report has explored the Olympic dataset through data cleaning, analysis, and visualization, uncovering key trends and relationships. Future work could delve deeper into specific sports or regions, employ predictive modeling to forecast outcomes or explore the impact of socio-economic factors on Olympic success.