Missing Data

Group No: 19

Vats Shah-202201417
Aniruddha-202411065
Monson-202411039

Date: September 19, 2024

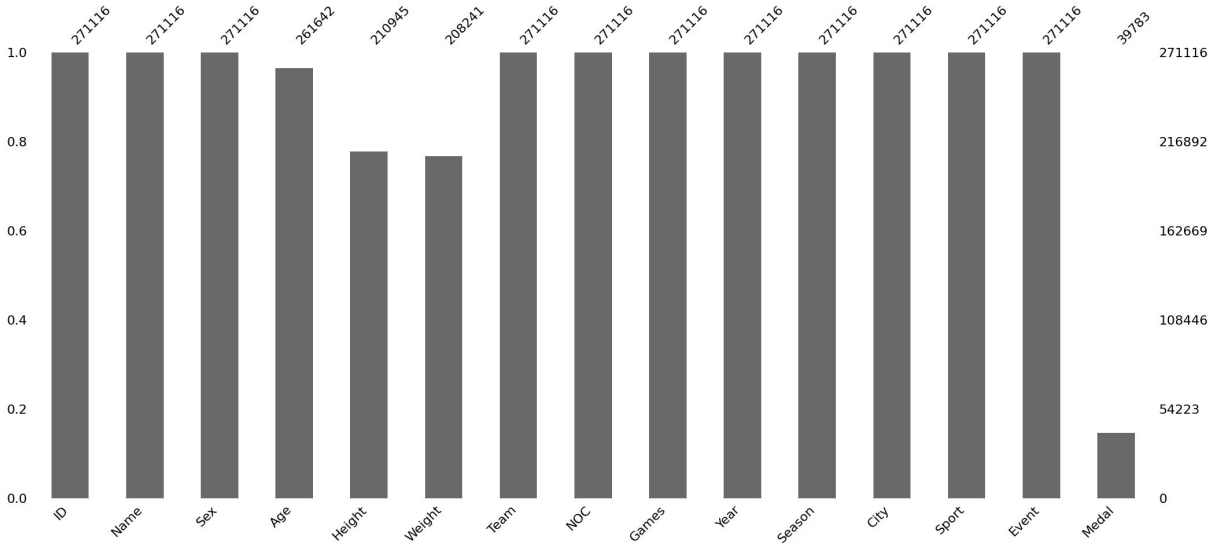
Barplot

The y-axis on the left side of the bar plot shows a scale from 0.0 to 1.0, where 1.0 represents complete data (i.e., 100% data completeness). A bar height lower than 1.0 suggests that missing values exist within the column.

On the right side, the axis represents index values, with the top-right corner indicating the total number of rows in the dataframe.

At the top of the plot, numbers display the count of non-null values for each column, giving a quick overview of data completeness.

- There are some values missing in column of Age and approximately about 25 percent missing values in columns of Height and weight and more than 80-90 percent values missing in column of Medal which is also consistent with our statistical findings



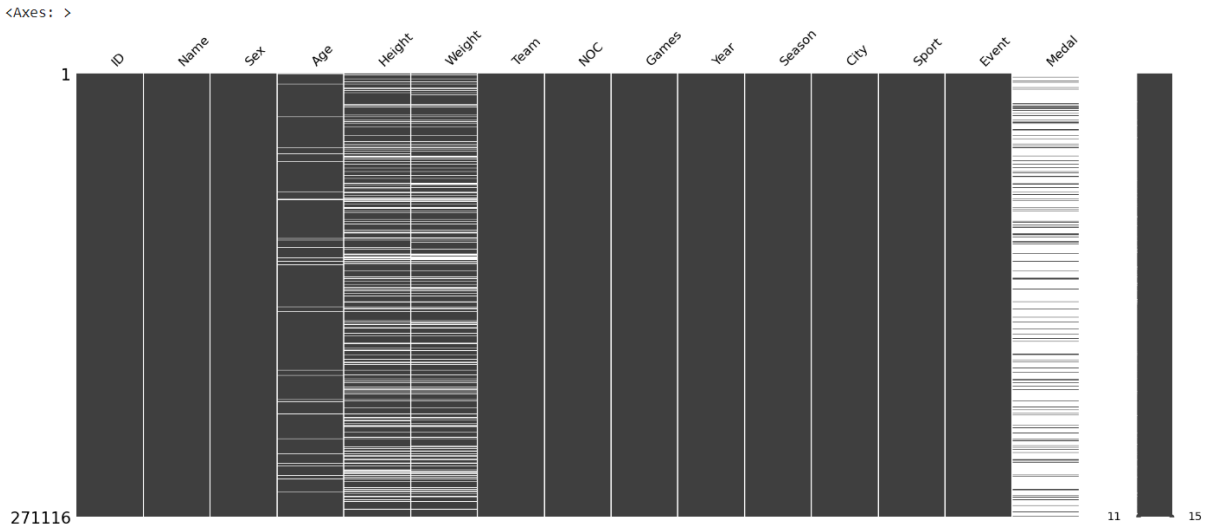
Matrixplot

The matrix plot is ideal for time-series or depth-related data. Each column is represented by a shaded color, indicating data presence (typically shaded grey) or absence (white). This makes it easy to spot missing data locations.

- White areas indicate missing values.
- The sparkline on the right shows the shape of data completeness and identifies the row with the fewest null values. At the bottom, you'll see the total number of columns.
- A row with no missing values will have a line to the far right, while an increase in missing values moves the line leftward.

To further analyze missing data, you can use matrix plots sorted by specific columns, such as gender.

- The white lines in certain columns (e.g., Age, Height, Weight, etc.) indicate missing data, while black areas indicate non-missing data. Columns like Medal and Age seem to have the most missing values, while columns such as ID, Name, NOC, and Sport appear to have very few or no missing values.



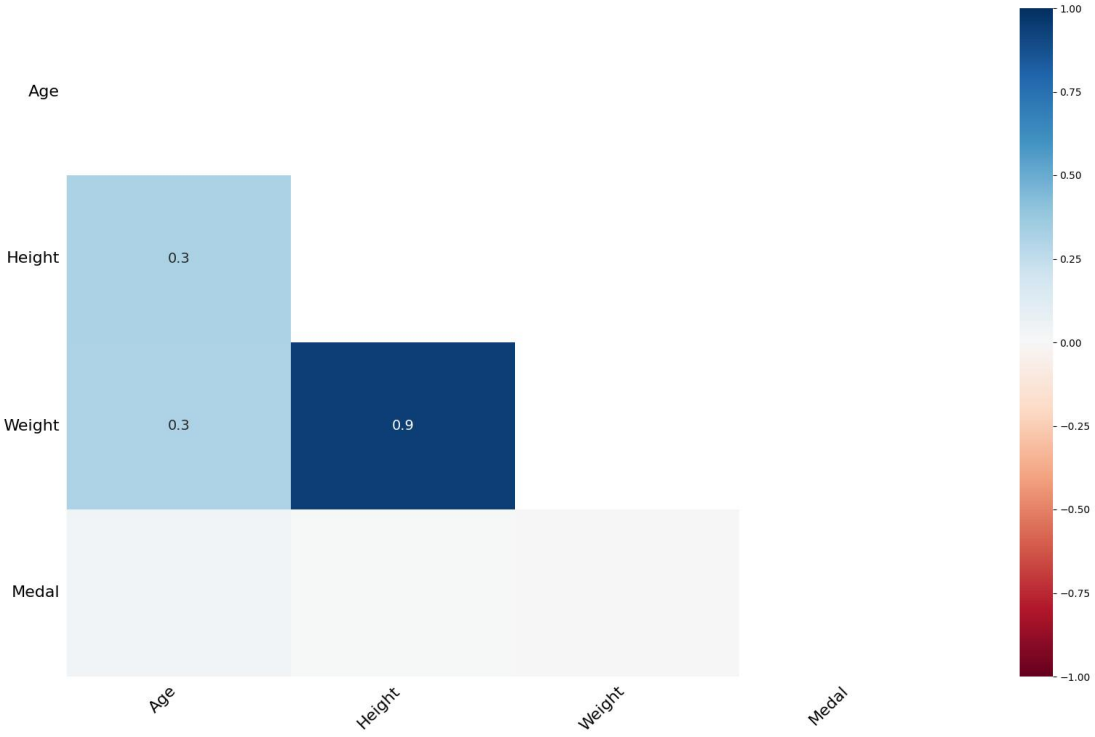
Heatmap

A heatmap helps identify correlations between missing values across columns. It highlights whether the presence of null values in one column relates to null values in another.

- Values close to 1 indicate a high correlation, meaning null values are present in both columns.
- Values close to -1 show an inverse correlation: one column has null values while the other does not.
- A value near 0 means there is little to no relationship between null values in the two columns.

For instance, a heatmap might reveal no strong correlation between the missingness in gender and other variables, suggesting that the data is Missing At Random (MAR).

- We can see that there is no correlation between the missingness in Age and Medal with rest of the variables

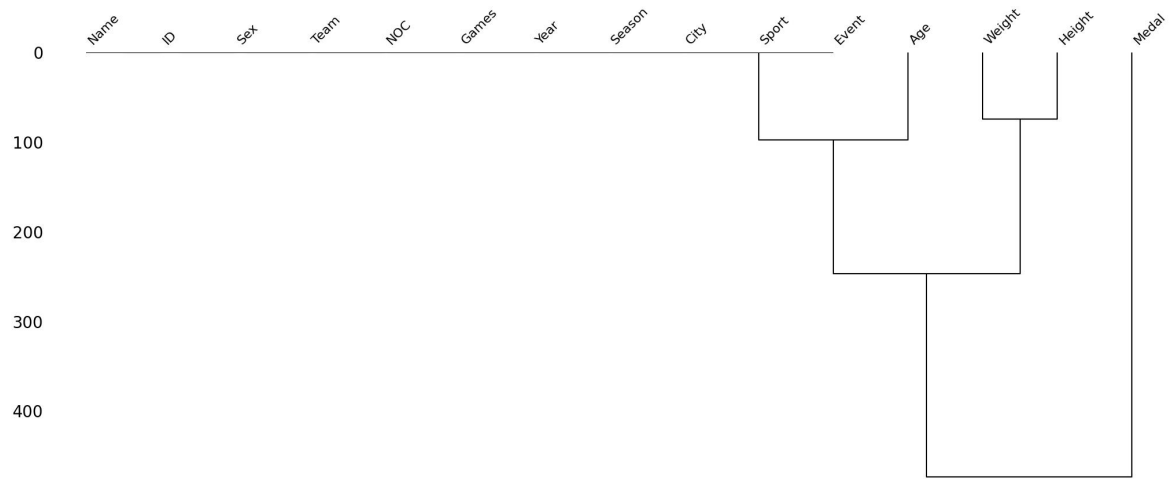


Dendrogram

The dendrogram is a tree-like graph created via hierarchical clustering. It groups together columns that have strong correlations in terms of missing data. Columns clustered together at level zero show direct relationships in missing values.

- Columns such as Age, Weight, and Height show significant clustering, indicating that they share similar missing data patterns.
- Other columns like Sport, Event, and Medal are also clustered together, suggesting related missingness, while Name, ID, and Sex have fewer missing values and appear more isolated.

This technique helps visualize relationships between missing values across the dataset.



Conclusion

Summary of Features in the Dataset

Complete Columns

Several columns have no missing values:

- **ID**
- **Name**
- **Sex**
- **Team**
- **NOC**
- **Games**
- **Year**
- **Season**
- **City**
- **Sport**
- **Event**

Columns with Missing Values

- **Age**: Missing **9,474** entries, which is **3.49%** of the data.
- **Height**: Missing **60,171** entries, accounting for **22.19%** of the data.
- **Weight**: Missing **62,875** entries, making up **23.19%** of the data.
- **Medal**: Missing **231,333** entries, which is the largest percentage of missing data at **85.33%**.

General Observation

Most demographic and event-related columns are complete, but physical measurements like **Age**, **Height**, and **Weight** have notable amounts of missing data. The **Medal** column has the highest percentage of missing values, as many participants may not have received medals.