**Dimension of dataset**: 3896 x 16

**What are the features in the dataset “schedules”?**

start\_date:

* Description: The start date and time of the event in the format YYYY-MM-DDTHH:MM.
* Purpose:: It marks when each specific sports event began.
* **Relation to Gender Disparity**: While these fields are primarily used to schedule events, they can reveal patterns of gender disparity if, for example, men’s events are scheduled at more favorable times (e.g., primetime slots), while women’s events may be scheduled at less favorable times.
* **Relation to Venue vs performance**: Understanding the start and end dates of events helps track venue utilization over time. It shows which venues are the busiest on certain days or during certain periods. This information is useful for scheduling, logistics, and ensuring that venues are not overbooked.

end\_date:

* Description: The end date and time of the event.
* Purpose: It indicates when each event concluded, giving the duration of the event.

day:

* Description: The date (without time) of the event in YYYY-MM-DD format.
* Purpose: Used to organize and filter events by day.
* **Relation to Gender Disparity**: The distribution of events for each gender across days may show patterns. For instance, if male events are scheduled more frequently on weekends (when more viewers are available) and female events on weekdays, that could indicate a form of scheduling bias.
* **Relation to Venue vs performance**: This helps break down the venue usage by each day, allowing for day-specific venue planning and optimization. High demand days at certain venues can reveal peak utilization times.

status:

* Description: The status of the event (e.g., "FINISHED").
* Purpose: Identifies whether an event is completed or ongoing, which could be useful in creating a status-based filter or grouping events.

discipline:

* Description: The name of the sport or event category (e.g., "Football", "Rugby Sevens").
* Purpose: Indicates which type of sport is being played, providing a classification for each event.
* **Relation to Gender Disparity**: Gender disparity can be examined by looking at the number of events for each discipline across genders. Certain sports may have more events or visibility for men compared to women or vice versa. For example, there may be a large number of football events for men but fewer for women.
* **Relation to Venue vs performance** : Different venues are likely specialized for specific sports. This allows for the analysis of which sports utilize which venues most frequently and how evenly distributed the disciplines are across venues.

discipline\_code:

* Description: Abbreviated code for the discipline (e.g., "FBL" for Football, "RU7" for Rugby Sevens).
* Purpose: A short code for easy reference to each sport or discipline.

event:

* Description: The specific gender category of the event (e.g., "Men").
* PurposeHelps identify which gender is participating in the event.
* **Relation to Gender Disparity**: The number of events associated with each gender can be counted and compared. If a certain gender has fewer events, it may indicate unequal representation or opportunity in that sport.

event\_medal:

* Description: A placeholder for the medal awarded (appears to have "0" for all entries in the screenshot).
* Purpose: Used to store the number of medals (gold, silver, bronze) .
* **Relation to Gender Disparity**: This feature helps to compare the number of medals won by each gender and gain respective insights on gender disparity.
* **Relation to Venue vs performance:** Medals won at each venue, Certain venues, like those dedicated to high-participation sports such as athletics and swimming, have a significantly higher count of medal-winning events compared to others.

phase:

* Description: The stage of the event (e.g., "Men's Group B").
* Purpose: Indicates which phase of the tournament or competition the event belongs to (e.g., group stages, semifinals).

gender:

* Description: The gender category of the participants (e.g., "M" for Men).
* Purpose: Helps classify the events by gender.
* **Relation to Gender Disparity**: By analyzing the distribution of events between males and females, you can identify if there are fewer events, opportunities, or resources allocated to one gender compared to the other.
* **Relation to Venue vs performance:**

event\_type:

* Description: Indicates the type of the event (e.g., "HTEAM", perhaps referring to a team event).
* PurposeSpecifies the format of the event (whether it’s a team-based competition or something else).

venue:

* Description: The full name of the venue where the event takes place (e.g., "Geoffroy-Guicha Stadium").
* Purpose: Shows the exact location where the event occurs, useful for logistics or filtering by location.
* **Relation to Gender Disparity**: You can investigate if one gender’s events are more often scheduled in prestigious or larger venues, while the other gender’s events are held in less prominent venues. This could indicate a disparity in how the events are valued and supported.

venue\_code:

* Description: A short code representing the venue (e.g., "STE" for Geoffroy-Guicha).
* Purpose: Abbreviation for quick identification of venues.

location\_description:

* Description: A description of the venue’s location.
* PurposeProvides additional context or address details about where the event is held.

location\_code:

* Description: A short code for the location (appears to repeat the venue\_code in the screenshot).
* Purpose: Similar to venue\_code, this gives a shorthand for identifying the venue location.

url:

* Description: A URL (web link) pointing to results or additional information about the event.
* Purpose: Directs users to external resources, likely where they can view event details or results.

### **How the Columns Relate to Each Other:**

* **Time & Status**: The start\_date, end\_date, and day columns are essential for understanding the scheduling of events. They pair with the status column to provide a timeline of events and their completion status.
* **Sport Information**: The discipline, discipline\_code, event, phase, and event\_type columns work together to classify the events by sport, gender, and the specific phase of the competition.
* **Venue Information**: The venue, venue\_code, location\_description, and location\_code columns describe where the events are held, making it easy to filter or sort events by location.
* **Additional Resources**: The url column provides external information, likely linking to more details or live results for each event.

**Statistics:**

There is only one feature with appropriate numeric data to find the mean ‘event\_medals’ - number of medals won by each discipline

| **mean** | 0.213864 |
| --- | --- |
| **std** | 0.653475 |
| **min** | 0.000000 |
| **25%** | 0.000000 |
| **50%** | 0.000000 |
| **75%** | 0.000000 |
| **max** | 3.000000 |

### **Insights:**

* The data is heavily skewed toward events where no medals were awarded, as indicated by the 25%, 50%, and 75% percentiles all being 0.
* Only a small fraction of events resulted in medals being awarded, which can be inferred from the mean of 0.21.
* The few events that do award medals might do so generously, as seen by the high standard deviation and the maximum value of 3 medals.

Possible Follow-Up:You might want to investigate which specific disciplines or events account for the awarded medals, as this could reveal trends in which sports are more competitive or likely to result in multiple medalists.

**Nominal or categorical data:** Using the function dtypes itself we can find out, here in this dataset only one feature called ‘event\_medals’ is numeric rest all are categorical.

**Missing values:** The 2 main columns focused in this dataset are most importantly ‘gender’ and then ‘venue’ as only these 2 relate to my problem statement of gender disparity and host vs performance.

gender 2

discipline discipline\_code event event\_medal phase gender \

49 Opening Ceremony CER NaN 0 NaN NaN

3894 Closing Ceremony CER NaN 0 NaN NaN

Insights: number of missing values in the feature ‘gender’ is 2 and on analyzing more those 2 rows is not relevent as it is talking about the opening ceremony and closing ceremony and these 2 dont have a gender and this 2 rows needs to be cleaned out as not relevenT.

venue 2

Discipline discipline\_code event event\_medal phase gender venue

49 Opening Ceremony CER NaN 0 NaN NaN NaN

3894 Closing Ceremony CER NaN 0 NaN NaN NaN

Insights : Number of missing values in the feature ‘gender’ is 2 and on analyzing more. those 2 rows is not relevant and doesn't affect my end goal, as it is talking about the opening ceremony and closing ceremony and these 2 don't have a gender.These 2 rows need to be cleaned out as not relevant.

**Possible Visualizations from this dataset:**

**Important visualization related to our final problem statement:**

### **Event Count by Gender and Discipline**

### **Columns: gender, discipline, event**

### **Visualization:** A stacked bar chart showing the number of events for each sport (discipline), separated by gender**.**

### **Insight:** This allows you to directly compare the number of events allocated to male and female participants in each sport. Gender disparity is evident if some disciplines heavily favor one gender**.**

### Insight we can get that relate to Gender Disparity

1. **Balanced Representation in Many Disciplines**: Many disciplines, such as athletics, swimming, judo, and fencing, now feature a similar number of events for male and female athletes. This marks progress toward gender equality compared to the 1980s, when events were largely male-dominated.
2. **Mixed-Gender Events on the Rise**: The inclusion of mixed-gender events (X) across disciplines like athletics and swimming highlights a growing focus on gender-inclusive formats, a notable change from 1986 when such events were rare. This shift promotes collaboration between male and female athletes in competition.

### 

### **Total Event Count by Gender**

### **Columns: gender, event**

### **Visualization:** A pie chart or bar chart showing the total number of events for males vs. females across all sports.

### **Insight:** This gives a big-picture view of whether there are more events for one gender overall, indicating general gender disparity.

### Insight we can get that relate to Gender Disparity

1. **Balance Between M and W**: The distribution between **M (49.8%)** and **W (46.0%)** is nearly balanced, indicating an almost equal participation rate among these two groups. This balance could suggest successful efforts towards gender inclusivity, especially if there was a significant gender disparity in the past.
2. **Significant Minority Representation (X and O)**: **X (3.6%)** and **O (0.5%)** represent non-binary or other gender categories. While these numbers are relatively small, they show that the event is inclusive of gender minorities, which could be an important consideration in fostering a diverse community. The presence of this representation, even though very negligible, could indicate progress towards broader gender inclusion, especially if these categories were absent or even smaller in earlier stages.

### 

### **Venue Usage by Gender**

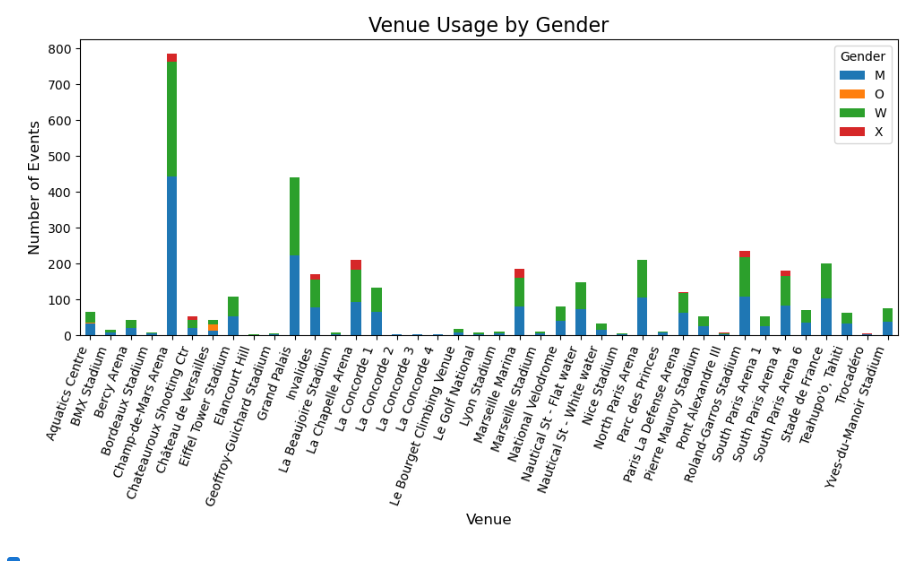
### **Columns: gender, venue, event**

### **Visualization:** A stacked bar chart or heatmap showing the distribution of male and female events across different venues.

### **Insight**: This helps to identify if certain venues are predominantly used for male or female events, which could indicate gender disparity in the allocation of resources (e.g., prestigious or larger venues going to one gender).

### Insight we can get that relate to Gender Disparity

1. Most venues like **Le Golf National**, **Pierre Mauroy Stadium**, and **La Chapelle Arena** show a more balanced representation of **M** and **W**, indicating progress in gender inclusivity. This balance may reflect efforts in these locations to bridge the gender gap, as both M and W seem to have close numbers of events.
2. The representation of **W (green)** in many venues has grown significantly, reducing the disparity that might have existed at the beginning. While there are still venues where **M (blue)** dominates, the balance at several venues indicates efforts to promote inclusivity.
3. Venues like **La Défense Arena** and **South Paris Arena 3** have events that include more representation for **X** and **O**, suggesting an improvement in the visibility and participation of gender minorities over time. Pointing towards increasing gender equity in event participation.

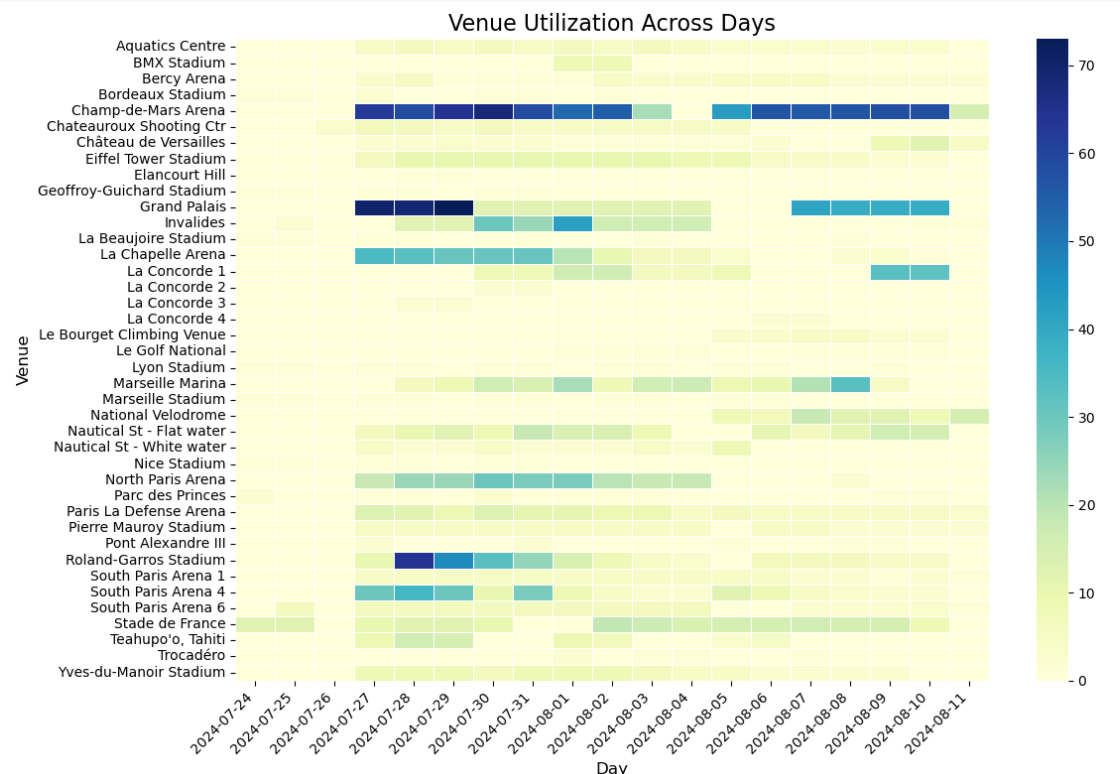


Showing a aggregate visualization for better readability

### 

### **Venue Utilization**

* **Columns**: venue, venue\_code, discipline, start\_date, end\_date
* **Visualization**: Heatmap or bar chart showing how frequently each venue is used across different days.
* **Insight**: Useful to understand which venues are most active and on which days.



### 

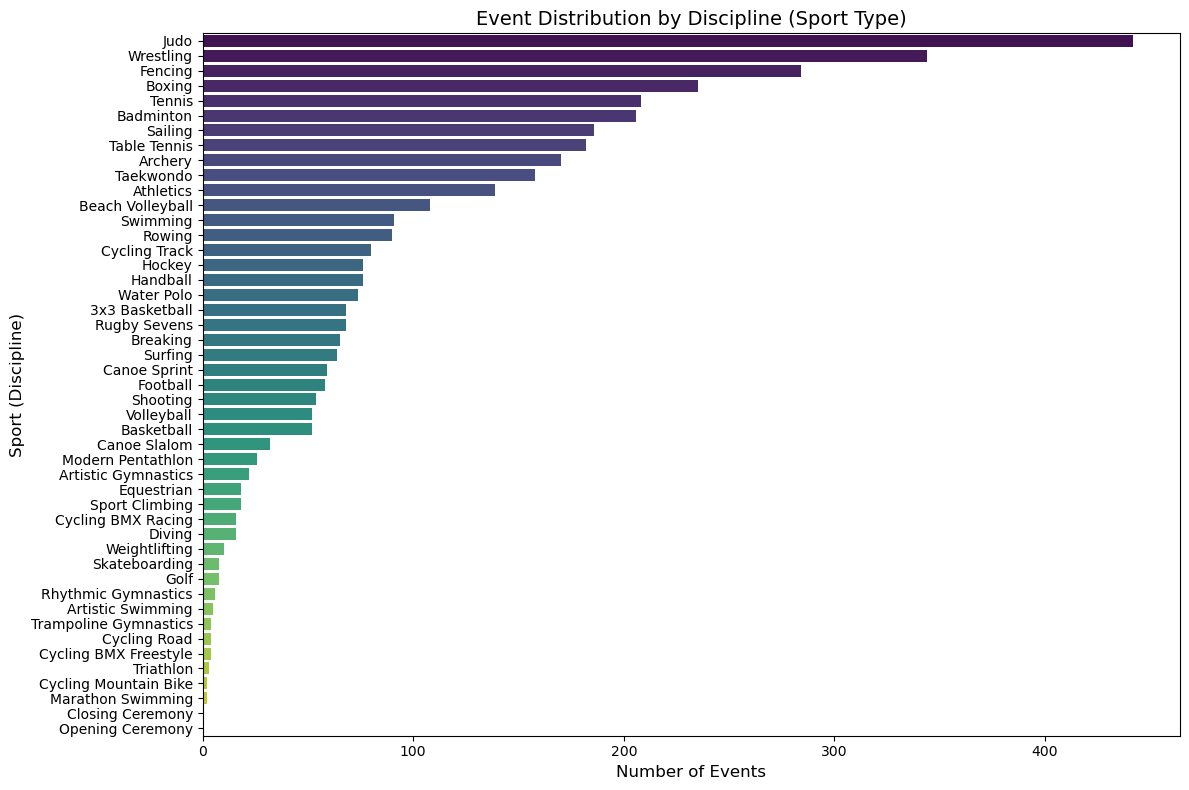
**Limitations:** Inoder to compare with previous datasets, to prove that gender disparity has gone down over the years, we don't have a feature in this dataset that shows the previous olympic data. This is a limitation of this data set. And the need to merge different datasets to conclude that gender disparity has gone down, which is our aim from this project.

THIS PARTICULAR FEATURE OF MEDAL And venue doesn't give us the insight we require to conclude to the final problem statement of host vs performance hence this limitation leads us to merge features across different datasets to align to the goal.

### **OTHER VISUALIZATIONS THAT ARE NOT RELATED TO THE PROBLEM STATEMENT BUT CAN DERIVE FROM THIS DATASET**

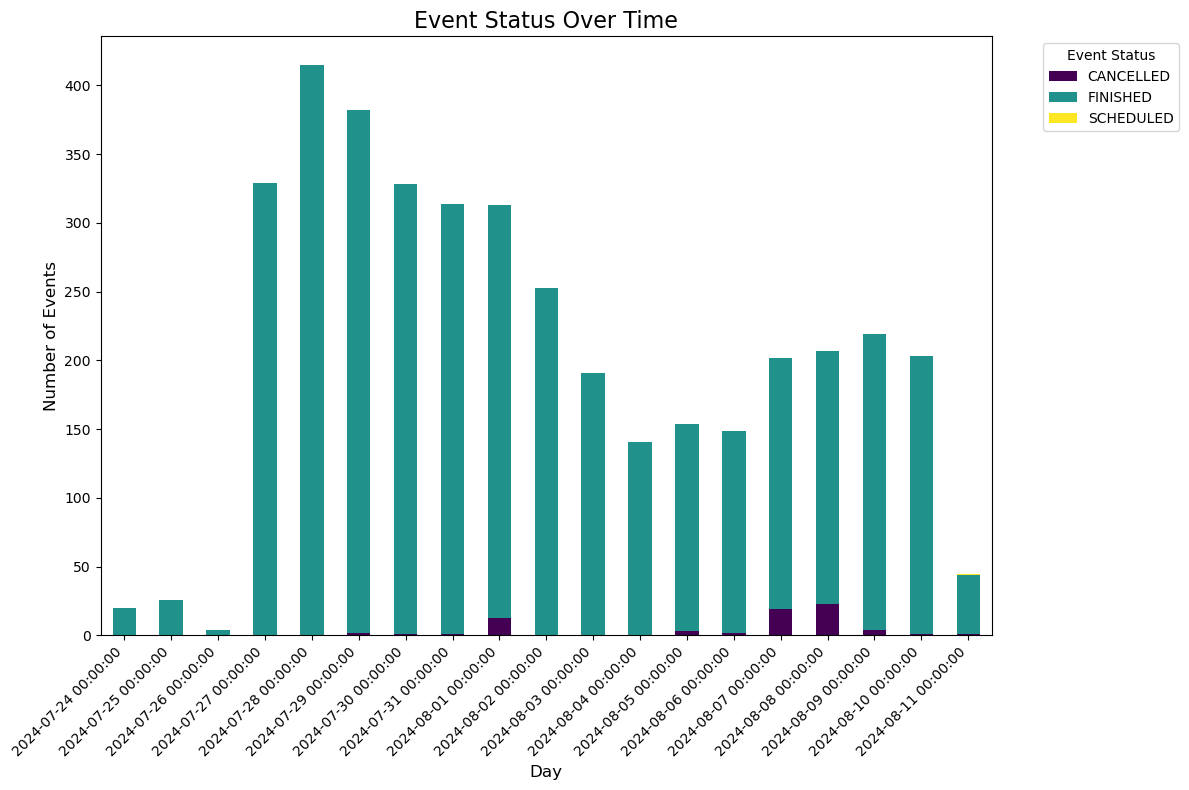
### **Event Distribution by Discipline (Sport Type)**

* **Columns**: discipline, discipline\_code, event, day
* **Visualization**: Bar chart or pie chart showing the number of events for each sport (e.g., Football, Rugby Sevens).
* **Insight**: This helps to see which sports have the most events in the Olympics.
* Sign



### **Event Status Over Time**

* **Columns**: day, status, start\_date, end\_date
* **Visualization**: Time series or stacked bar chart showing the number of events per day, segmented by their status (e.g., "FINISHED").
* **Insight**: This will show the timeline of events over the course of the Olympics and how many events are ongoing or finished at different times.



### **Phase Progression of Events**

* **Columns**: discipline, phase, event\_type, day
* **Visualization**: A timeline chart or Sankey diagram showing the progression of events across phases (e.g., from group stages to finals).
* **Insight**: Tracks how events progress through different competition stages.

### **Event Duration Analysis**

* **Columns**: start\_date, end\_date, discipline
* **Visualization**: Box plot or histogram showing the duration of events, broken down by discipline.
* **Insight**: Helps compare how long events from different sports last, allowing for insights into the scheduling needs of each discipline.

### **Location-Wise Event Distribution**

* **Columns**: venue, venue\_code, discipline, location\_code
* **Visualization**: Geographical map with venue locations, color-coded by event count or discipline.
* **Insight**: Provides a geographical overview of where events are taking place.

### **Event Count by Day**

* **Columns**: day, discipline, status
* **Visualization**: Line or area chart showing how many events are scheduled per day.
* **Insight**: A good high-level view to see how event activity varies over the course of the Olympics.

### **Medal Event Analysis**

* **Columns**: event\_medal, discipline, event\_type, phase
* **Visualization**: Bar chart showing how many medal events occur for each discipline or event type.
* **Insight**: Although not populated in the current dataset, once filled, it can show which sports award the most medals and at what stages.

### **Events by Type (Team vs. Individual)**

* **Columns**: event\_type, discipline, gender, phase
* **Visualization**: Pie chart or bar chart showing the split between team and individual events for each sport.
* **Insight**: Understand the prevalence of team vs. individual competitions in the Olympics.

DATA DISTRIBUTION:

### **start\_date & end\_date**

* **Expected Distribution**:
  + The event times would likely follow a uniform distribution over the course of the Olympics, assuming events are evenly scheduled throughout the day. However, more events may cluster around prime viewing hours (e.g., afternoon and evening).
  + Most events would start and end within a few hours. If the Olympics last for several weeks, events will span across all days.
* **Potential Visualization**: Histogram or line chart showing the frequency of event start times and durations over the days.

### **2. day**

* **Expected Distribution**:
  + This column would likely have a uniform distribution over the span of the Olympic Games. Each day should have a similar number of events, with perhaps a slight increase in event counts on weekends or as the Games progress toward the finals.
* **Potential Visualization**: Bar chart of events per day.

### **3. status**

* **Expected Distribution**:
  + Initially, most events will have a "Scheduled" or "Upcoming" status, and as time passes, more events will switch to "Finished". Toward the end of the dataset, almost all events would have a "Finished" status.
* **Potential Visualization**: A time series chart showing the proportion of finished, ongoing, and scheduled events over time.

### **4. discipline**

* **Expected Distribution**:
  + The distribution of sports disciplines might be skewed. High-demand sports like Football, Track & Field, or Swimming could have more events, while niche sports might have fewer events.
* **Potential Visualization**: A bar chart or pie chart showing the count of events for each sport.

### **5. discipline\_code**

* **Expected Distribution**:
  + This would mirror the discipline column but in abbreviated form. The frequency of each discipline code would be similar to the distribution of discipline.
* **Potential Visualization**: Similar to discipline, bar chart or pie chart.

### **6. event**

* **Expected Distribution**:
  + If this column refers to gender (as it does in the image, showing "Men"), we might expect a relatively equal distribution between "Men" and "Women" events, depending on how evenly distributed sports are across genders. In some disciplines, there may be more men's events, and in others, more women's.
* **Potential Visualization**: A bar chart showing the split between men's and women's events.

### **7. event\_medal**

* **Expected Distribution**:
  + Initially, most events would have no medals assigned (represented as 0), but as events finish and medals are awarded, the distribution will become more varied, with values representing gold, silver, and bronze.
* **Potential Visualization**: Stacked bar chart showing the progression of medal assignments over time.

### **8. phase**

* **Expected Distribution**:
  + For team sports (e.g., football, rugby), earlier phases (group stages) will have many events, while later phases (like finals) will have fewer events. This leads to a pyramid-like distribution, where the number of events decreases as the competition progresses.
* **Potential Visualization**: A pyramid chart showing the number of events per phase or a Sankey diagram for event progression.

### **9. gender**

* **Expected Distribution**:
  + Similar to the event column, we expect a relatively equal distribution between "M" (Men) and "W" (Women), with possible slight variations based on the sports or competitions.
* **Potential Visualization**: Bar chart showing the breakdown of events by gender.

### **10. event\_type**

* **Expected Distribution**:
  + The distribution here will depend on the types of events in the dataset (e.g., team vs. individual). For team sports, expect "HTEAM" (if this represents team-based events) to be more prevalent.
* **Potential Visualization**: Pie chart or bar chart to show the distribution of team vs. individual events.

### **11. venue & venue\_code**

* **Expected Distribution**:
  + Some major stadiums or venues (e.g., Stade de France, Parc des Princes) may host many events and therefore have a higher frequency, while smaller or more specialized venues may host fewer events.
  + The distribution may be right-skewed, where a few large venues host many events, and many smaller venues host fewer events.
* **Potential Visualization**: Bar chart showing the number of events at each venue.

### **12. location\_description & location\_code**

* **Expected Distribution**:
  + This would follow a similar pattern to venue. Some locations (like large cities or regions) will host a higher number of events, while other locations may have fewer.
* **Potential Visualization**: Map-based visualization or a bar chart showing the distribution of events across locations.

### **13. url**

* **Expected Distribution**:
  + Every row should have a unique URL for event results or information. The distribution wouldn't be directly visualized but would be important for linking events to external resources.
* **Potential Visualization**: Not typically visualized, but used for linking out to event results.

To do

* Limitation of dataset and problem and if significance to our goal or not —> insights part