Codedex:
Summer Hackathon

Track 3
Predict 2024 Olympics Champion

Team Members:
Ahmed Alyafai
Amiraj Singh
Jolene Tan

Table of Contents:

1.0 Introduction 2

2.0 Assumption 2

3.0 Graphs and Explanation 3-4

3.1 Teams Dataset 3-4

4.0 Countries and Disciplines 5-6

4.1 Athletes Dataset 6-8

5.0 Medals Won by each Country 8-12

6.0 Conclusion 12-13

7.0 References 13

1.0 Introduction:

The Olympic Games are the apex of international athletic competition, attracting competitors from all over the world to display their prowess and pursue perfection. We are using data science as part of our hackathon to examine the lengthy history of Olympic performances and forecast who might win in 2024.

Our collection includes a wealth of data about Olympic athletes, teams, coaches, and the medals that each nation has earned over the years. Our goal in examining these data sets is to find trends and information that can guide our forecasts. This study shows how data science techniques may be applied practically in real-world circumstances, while also highlighting the power of data analysis in the sports industry.

Through meticulous data cleaning, exploratory data analysis, and the application of predictive modeling, we will identify key factors that contribute to Olympic success. Our goal is to provide a comprehensive and accurate forecast of the countries most likely to excel in the upcoming Olympic Games, offering a glimpse into the future of international sports competition.

2.0 Assumptions:

We assume that historical performance is a strong predictor of future success, meaning past medal counts and rankings can inform our predictions for the 2024 Olympics. The dataset provided is assumed to be complete and accurate, encompassing all necessary information on athletes, coaches, teams, and medals. We also presume that the performance of athletes and teams will remain relatively consistent, without major disruptions such as injuries, retirements, or significant improvements that could alter the expected outcomes. Lastly, we rely on the idea that external factors such as changes in training regimes, political influences, or economic conditions have minimal impact on the athletic performance analyzed in this dataset.

OLYMPICS 2024 POTENTIAL TOP 5 WINNERS :

- UNITED STATES OF AMERICA
- REPUBLIC OF CHINA
- RUSSIAN OLYMPIC COMMITTEE
- GREAT BRITAIN
- JAPAN

Read below to find out how we obtained this result

3.0 Graphs and Code:

3.1 Teams Dataset

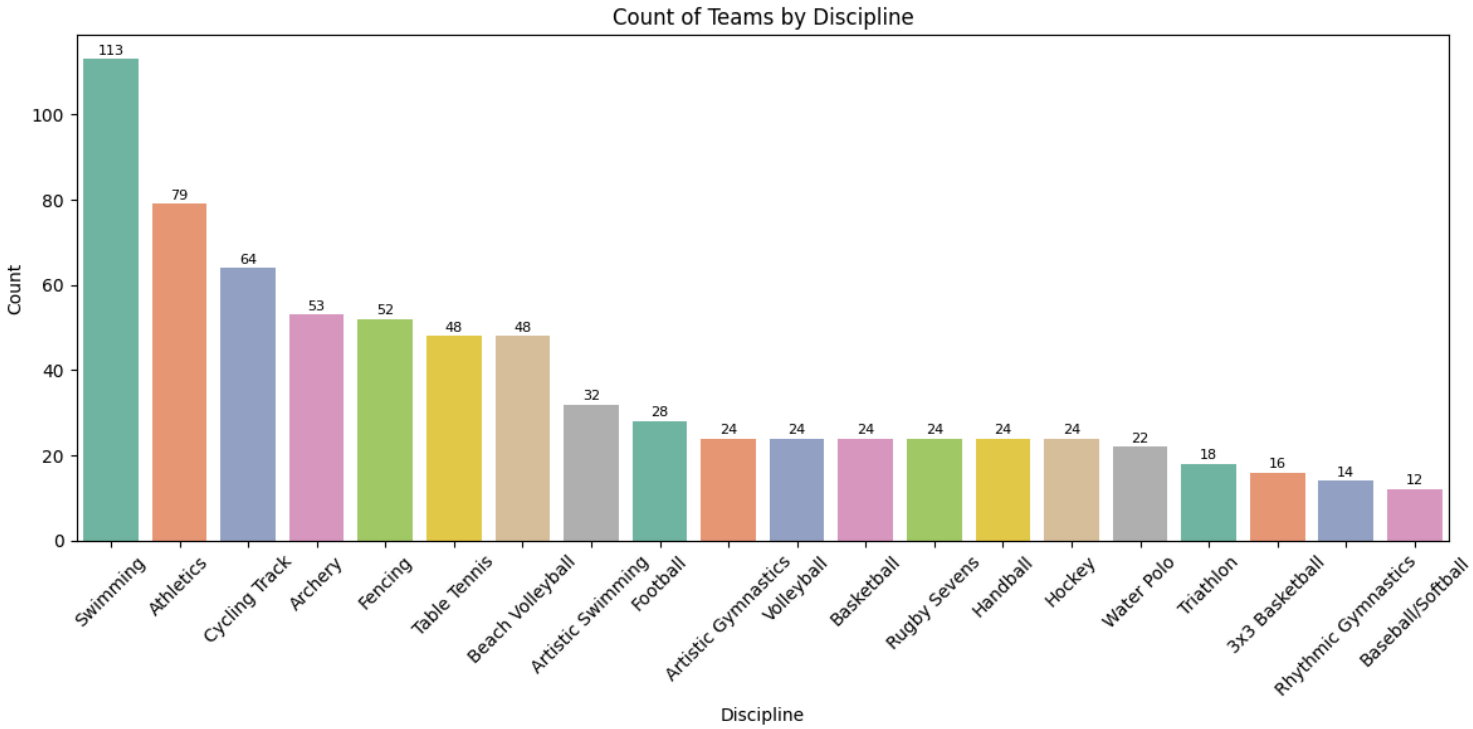

Figure 1.1 - Bar Graph of the Team Dataset

We have taken the dataset from an online source and have analyzed the data using Python Pandas. From the graph above, it can be said that Swimming has the most number of teams participating therefore higher numbers of chances to win medals are in the Swimming category. Then we would have to take a look at the country with the most teams participating in the Swimming Category which would give them the upper hand. The x-axis represents the disciplines, and the y-axis represents the count of teams. The chart title is "Count of Teams by Discipline."

Code

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
import seaborn as sns
file_path = r'C:\Users\amiraj\Downloads\Teams.xlsx'
df = pd.read_excel(file_path)


print(df.head())
print(df.columns)


print(df.dtypes)
print(df.isnull().sum())
teams_by_discipline = df['Discipline'].value_counts()
teams_by_event = df['Event'].value_counts()
plt.figure(figsize=(12, 6))
sns.countplot(x='Discipline', data=df, palette='Set2', order=teams_by_discipline.index)
plt.title('Count of Teams by Discipline')
plt.xlabel('Discipline')
plt.ylabel('Count')
plt.xticks(rotation=45)


for index, value in enumerate(teams_by_discipline):
        plt.text(index, value + 0.5, str(value), ha='center', va='bottom', fontsize=8)

plt.tight_layout()
plt.show()


plt.figure(figsize=(12, 6))
sns.countplot(x='Event', data=df, palette='Set3', order=teams_by_event.index)
plt.title('Count of Teams by Event')
plt.xlabel('Event')
plt.ylabel('Count')
plt.xticks(rotation=45)


for index, value in enumerate(teams_by_event):
        plt.text(index, value + 0.5, str(value), ha='center', va='bottom', fontsize=8)

plt.tight_layout()
plt.show()
```

4.0 Countries and Disciplines (Events) Dataset

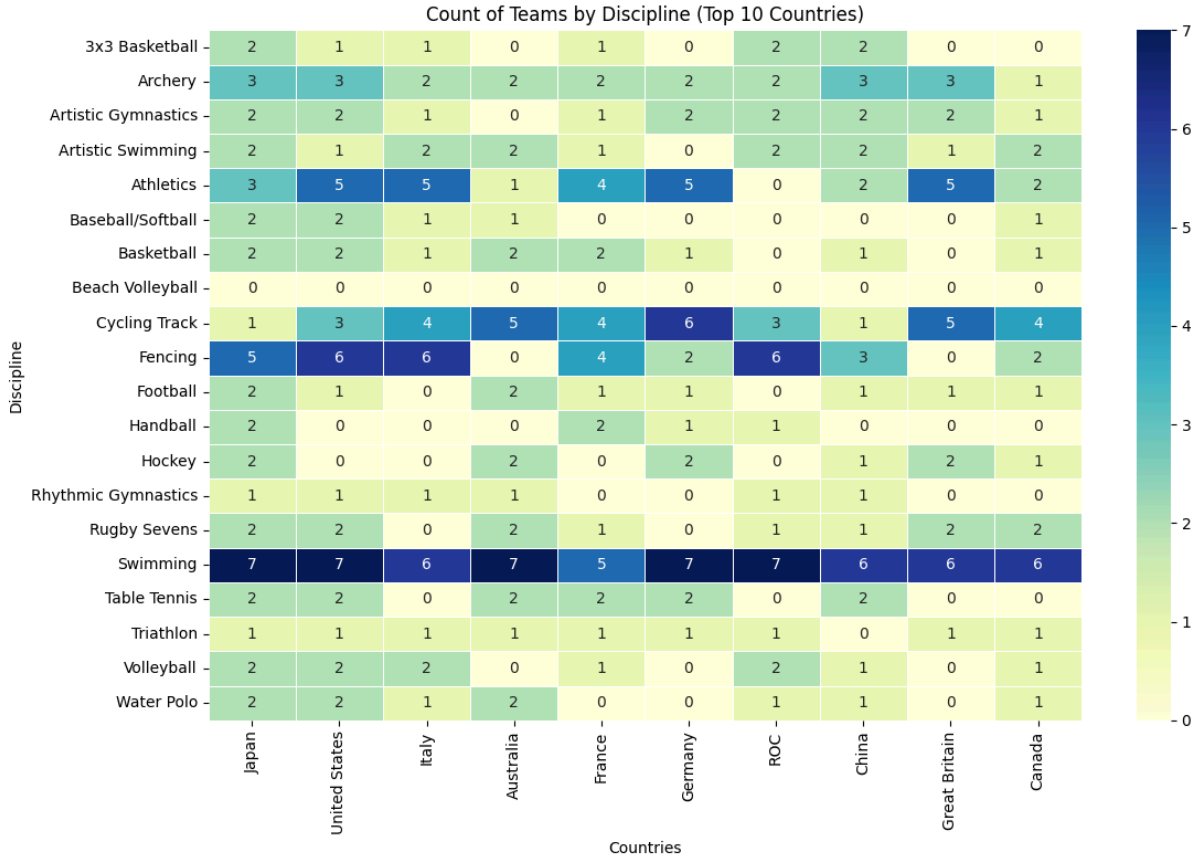

Figure 1.2 - Heatmap Dataset

As we can see here, Swimming has the most number of participants with Japan, United States of America, Australia, Germany, Republic of China leading with 7 each. Looking at the other events, the United States of America leads with the most participants. Therefore, giving them an advantage to winning more medals. The x-axis represents the countries, and the y-axis represents the disciplines. The color intensity increases with the count, with a scale ranging from 0 (light yellow) to 7 (dark blue).

Code:

```
import pandas as pd
 import matplotlib.pyplot as plt
 import seaborn as sns

 file_path = r'C:\Users\amiraj\Downloads\Teams.xlsx'
```

```
 team_discipline_counts = df.groupby(['Discipline',
'Name']).size().reset_index(name='Count')


 team_discipline_pivot = team_discipline_counts.pivot(index='Discipline',
columns='Name', values='Count').fillna(0)
 country_counts = team_discipline_pivot.sum(axis=0)
 top_10_countries = country_counts.nlargest(10).index
 team_discipline_pivot_top10 = team_discipline_pivot[top_10_countries]
 plt.figure(figsize=(12, 8))
 heatmap = sns.heatmap(team_discipline_pivot_top10, annot=True, fmt='.0f',
cmap='YlGnBu', linewidths=.5)

 plt.title('Count of Teams by Discipline (Top 10 Countries)')
 plt.xlabel('Countries')  # Adjusted xlabel here
 plt.ylabel('Discipline')
 plt.xticks(rotation=90)
 plt.yticks(rotation=0)
 plt.tight_layout()
 plt.show()
```

4.1 Athletes Dataset

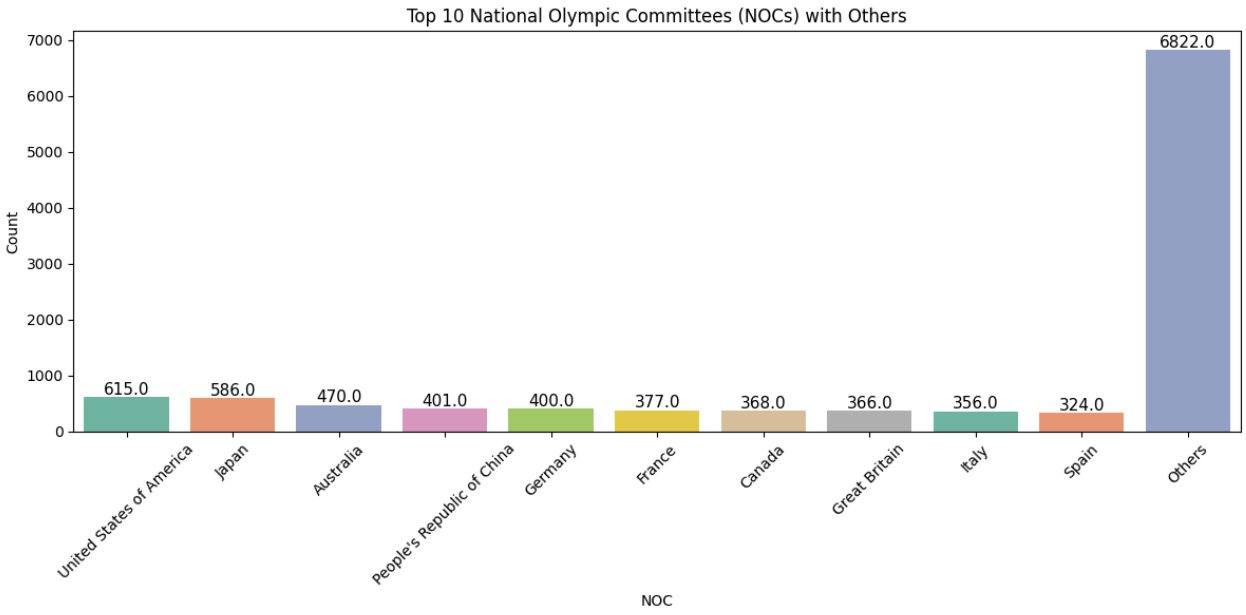

Figure 1.3 - Bar Graph of Athletes Dataset

This bar chart shows the count of National Olympic Committees (NOCs) for the top 10 countries and an "Others" category which represents the count of NOCs not included in the top 10. The counts for each country are as follows:

- United States of America: 615.0
- Japan: 586.0
- Australia: 470.0
- People's Republic of China: 401.0
- Germany: 400.0
- France: 377.0
- Canada: 368.0
- Great Britain: 366.0
- Italy: 356.0
- Spain: 324.0

The "Others" category has a significantly higher count of 6822.0. The x-axis represents the NOCs by country, and the y-axis represents the count of these committees. The countries are labeled and color-coded, with the bars showing the count for each respective NOC. The "Others" category has a bar that towers over the rest, indicating a much higher count compared to any individual country. The chart title is "Top 10 National Olympic Committees (NOCs) with Others."

Code

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file_path = r'C:\Users\amiraj\Downloads\Athletes.xlsx'
df = pd.read_excel(file_path)

nocs_counts = df['NOC'].value_counts()

top_n = 10
top_nocs = nocs_counts.head(top_n)
other_nocs_count = nocs_counts.iloc[top_n:].sum()

df['NOC'] = df['NOC'].apply(lambda x: x if x in top_nocs else 'Others')

df['NOC'] = pd.Categorical(df['NOC'], categories=list(top_nocs.index) +
['Others'], ordered=True)

plt.figure(figsize=(12, 6))
ax = sns.countplot(x='NOC', data=df, palette='Set2')

for p in ax.patches:
      ax.annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2.,
p.get_height()),
            ha='center', va='center', fontsize=11, color='black', xytext=(0,
5),
```

```
            textcoords='offset points')

plt.title(f'Top {top_n} National Olympic Committees (NOCs) with Others')
plt.xlabel('NOC')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

5.0 Medals Won by each Country (2021) Dataset

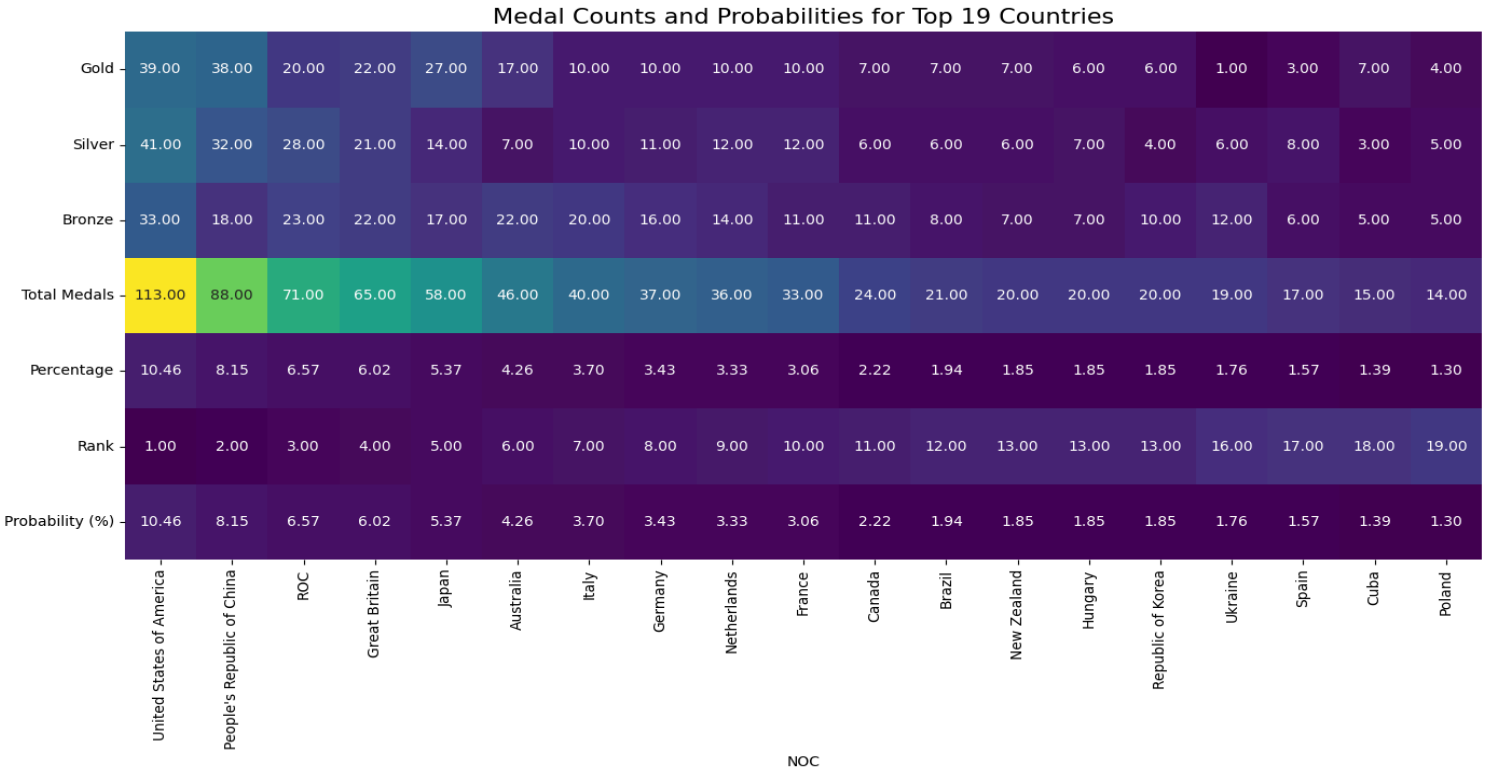

Figure 1.4 - Heatmap for Medals Won by Each Country with Total Medals ,Rank and Percentage

In Figure 1.4, we present the distribution of gold, silver, and bronze medals won by the top 10 countries based on their rankings. We then calculated the total number of medals each country won and ranked them accordingly. Using this ranking system, we determined the probability of each country winning the most medals. The results indicate that the United States has the highest probability at 10.46%, followed by the People's Republic of China at 8.15%, ROC at 6.57%, and so forth. The countries are ranked in descending order based on these probabilities, highlighting their likelihood of achieving the highest medal

count. The x-axis of the graph represents countries by their National Olympic Committees (NOCs) while the y-axis displays various metrics. Each cell in the heatmap shows the value of a specific metric for a given country, with the color indicating the magnitude of that value.

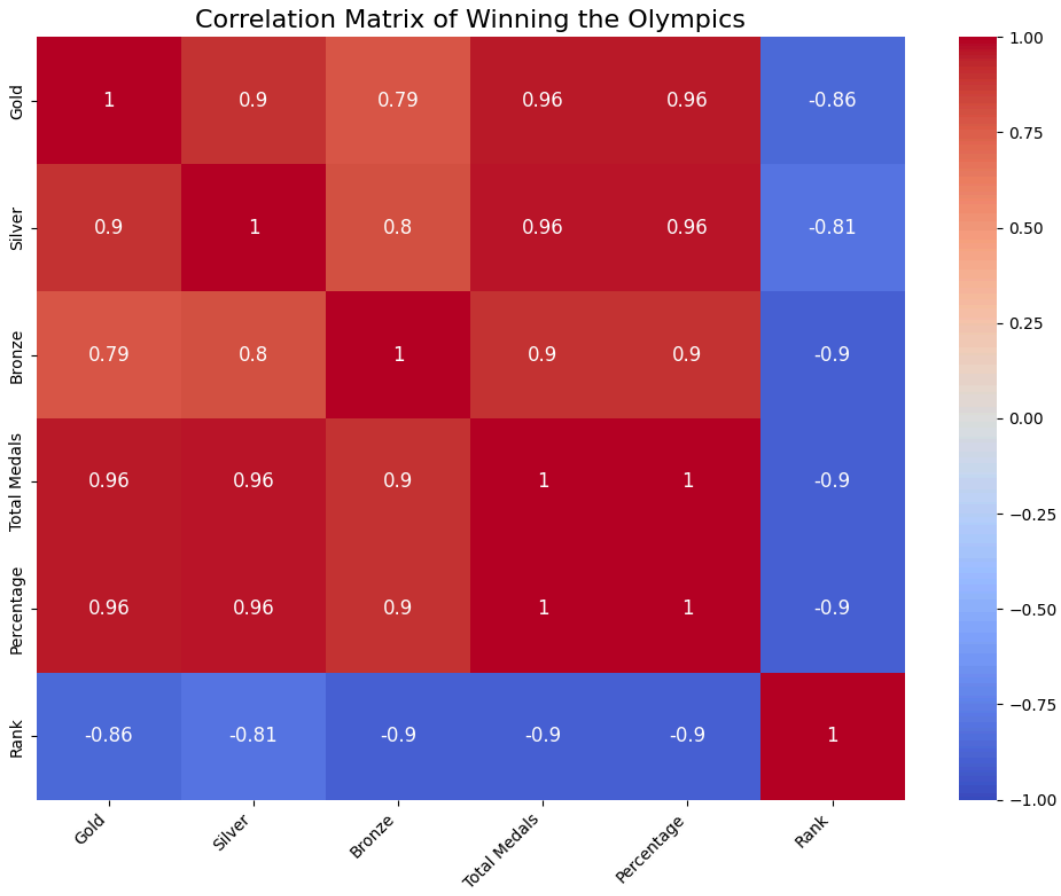

Figure 1.5 - Correlation Plot Medals

The image is a correlation matrix heatmap titled "Correlation Matrix of Winning the Olympics." It visualizes the correlation coefficients between various variables related to Olympic performance. The variables include Gold, Silver, Bronze medals, Total Medals, Percentage, and Rank.

Here's a breakdown of the heatmap:

1. Gold:
- Strong positive correlation with Silver (0.9), Bronze (0.79), Total Medals (0.96), and Percentage (0.96).
- Strong negative correlation with Rank (-0.86).

2. Silver:
- Strong positive correlation with Gold (0.9), Bronze (0.8), Total Medals (0.96), and Percentage (0.96).
- Strong negative correlation with Rank (-0.81).

3. Bronze:
- Strong positive correlation with Gold (0.79), Silver (0.8), Total Medals (0.9), and Percentage (0.9).
- Strong negative correlation with Rank (-0.9).

4. Total Medals:
- Strong positive correlation with Gold (0.96), Silver (0.96), Bronze (0.9), and Percentage (1.0).
- Strong negative correlation with Rank (-0.9).

5. Percentage:
- Strong positive correlation with Gold (0.96), Silver (0.96), Bronze (0.9), and Total Medals (1.0).
- Strong negative correlation with Rank (-0.9).

6. Rank:
- Strong negative correlation with Gold (-0.86), Silver (-0.81), Bronze (-0.9), Total Medals (-0.9), and Percentage (-0.9).

The color intensity represents the strength of the correlation, with red indicating positive correlations and blue indicating negative correlations. A value of 1.0 indicates a perfect positive correlation, while a value of -1.0 indicates a perfect negative correlation.

In summary, winning more Gold, Silver, and Bronze medals, as well as having a higher percentage of wins and total medals, is strongly correlated with a lower (better) rank in the Olympics.

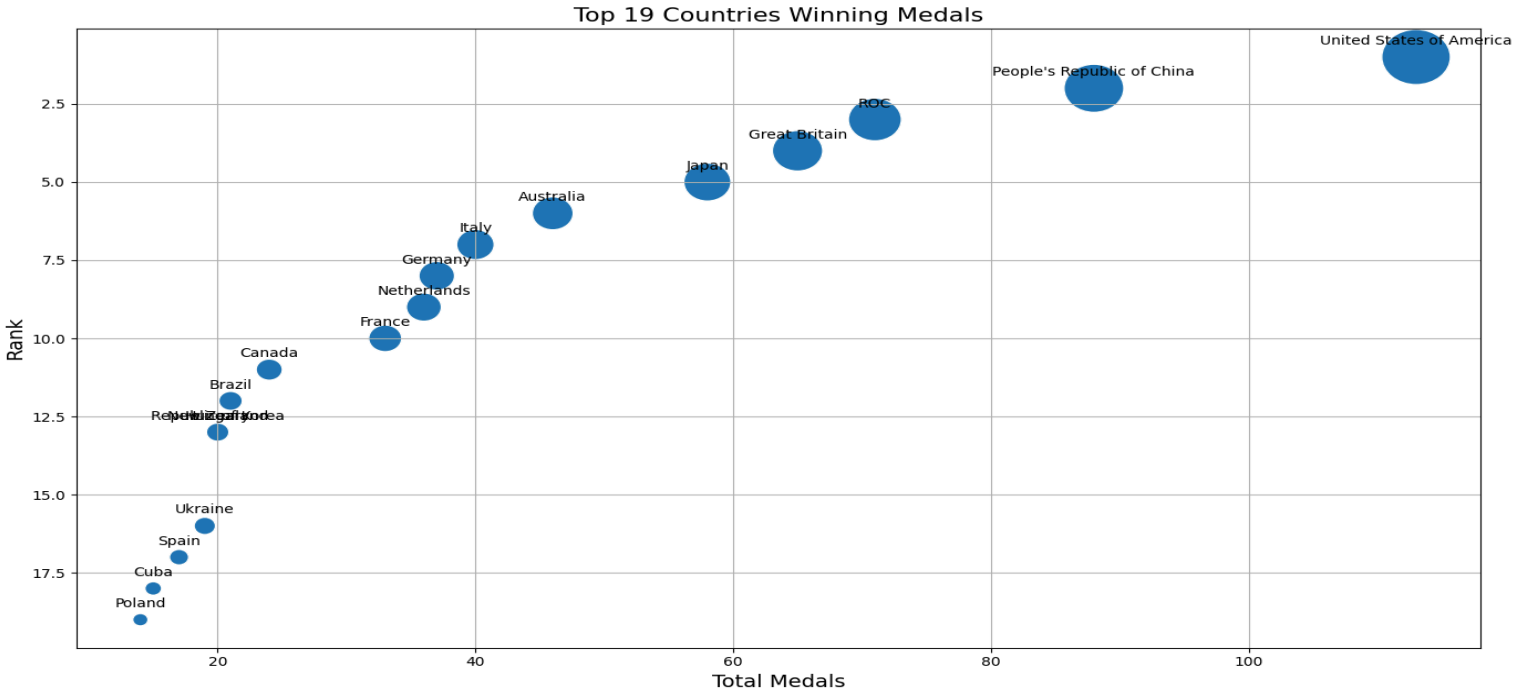

Figure 1.6 - Scatter Plot for the Number of Medals Won by Each Country

The image is a bubble chart titled "Top 19 Countries Winning Medals." It shows the relationship between the total number of medals won by each country and their rank in the Olympics. The x-axis represents the total number of medals, while the y-axis represents the rank, with a lower number indicating a better rank. The size of the bubbles represents an additional variable, possibly the number of gold medals or another significant metric.

Here's a detailed breakdown of the chart:

1. United States of America:
- Highest total medals (around 113).
- Rank is 1, indicating the best performance.
- Largest bubble, suggesting it has a high number of gold medals or a similar metric.

2. People's Republic of China:
- Second-highest total medals (around 85).
- Rank is around 2.5.
- Large bubbles, indicating a strong performance in key metrics.

3. ROC (Russian Olympic Committee):
- Total medals around 70.
- Rank is around 5.
- Relatively large bubble.

4. Great Britain:
- Total medals around 65.
- Rank is around 4.
- Moderately large bubble.

5. Japan:
- Total medals around 60.
- Rank is around 6.
- Moderate bubble size.

6. Australia:
- Total medals around 50.
- Rank is around 7.5.
- Moderate bubble size.

7. Italy, Germany, Netherlands, France:
- Total medals between 30-45.
- Ranks between 7.5 and 10.
- Moderate bubble sizes.

8. Canada, Brazil:
- Total medals around 25-30.
- Rank around 11.
- Smaller bubble sizes.

9. Other countries (Poland, Cuba, Spain, Ukraine, Hungary, Republic of Korea):
- Total medals between 15-25.
- Ranks between 12.5 and 18.
- Smaller bubble sizes.

The chart clearly shows that countries with more total medals tend to have better ranks (lower rank numbers). The bubble sizes provide an additional layer of information, highlighting countries with particularly strong performances in specific metrics (likely gold medals). The United States stands out as the top performer in terms of both total medals and rank.

6.0 Conclusion

In conclusion, leveraging Python along with essential libraries like pandas, matplotlib, seaborn, and sqlalchemy has proven invaluable in our data analysis and visualization endeavors. By working with datasets from the 2021 Olympics sourced from Kaggle, we meticulously cleaned

and processed each dataset. Given the vast scope of countries involved, we focused our analyses on the Top 10 countries, ensuring clarity and relevance in our graphical representations.

This structured approach not only facilitated comprehensive data management but also enabled us to derive actionable insights that could potentially inform and enhance preparations for the upcoming Olympics in 2024. By harnessing the power of these tools, we are better equipped to anticipate and interpret outcomes, contributing to a more informed strategy and decision-making process moving forward.

References:

2021 Olympics in Tokyo. (2021, August 17). Kaggle. https://www.kaggle.com/datasets/arjunprasadsarkhel/2021-olympics-in-tokyo?resource=download

Pandas-Dev. (n.d.). *GitHub - pandas-dev/pandas: Flexible and powerful data analysis / manipulation library for Python, providing labeled data structures similar to R data.frame objects, statistical functions, and much more*. GitHub. https://github.com/pandas-dev/pandas