Final Project - Phase 4-5

Collaborators: Jeremy Cortez (jac722), Rachel Bethke (rkb76)

Table of Contents:

1. Introduction

Olympic success has long been viewed as a matter of national pride, with countries investing heavily in athlete development, training facilities, and sports infrastructure. In our project, we wanted to consider how economic wealth translate directly into medals? Existing data seems to suggest that wealthier countries tend to win the majority of medal, due to factors such as their ability to fund extensive training programs, provide access to advanced sports science, and support larger teams of athletes. However, the relationship between economic resources and athletic success may be more nuanced than simple correlations suggest.

This project looks at how a country's Gross Domestic Product (GDP) relates to

Olympic performance across different types of sports during the 1992-2016 Summer and Winter Games. We investigate two complementary questions: first, whether countries at different economic levels specialize in different types of sports—specifically, whether lower-GDP nations earn a higher proportion of medals in less resource-intensive sports like archery and shooting compared to physically demanding sports like basketball and athletics. Second, we look at if physical attributes (i.e. height) are good predictors of performance to different degrees across differe t sport types, looking at if the biomechanical advantages of height matter more in physically demanding sports than in precision-based events.

Our analysis revealed complex relationships between economic resources and Olympic success:

1. GDP and total medals: Higher-GDP countries win more medals overall
2. No sport-type specialization: Lower-GDP countries do NOT earn disproportionately more medals in less physically demanding sports (archery, shooting) versus wealthy nations
3. Height matters more in physical sports: Biomechanical advantages predict success more strongly in physically demanding sports (Cohen's d=0.17) than precision sports (Cohen's d=0.10)

Conclusion: While economic resources correlate with overall Olympic success, the relationship is nuanced—wealth doesn't dictate specialization in specific sport categories, but physical attributes matter more where physicality is central.

Research Questions:

Does a Country's GDP have a significant impact on its Olympic performance? How does this relationship vary across different Olympic Games?

Does Physical Attributes (e.g. height) play a role in bettering the performance of athletes in non Physically demanding sports (e.g. Archery)?

Why these questions matter:

These questions let us test both macro-level (country economic strength) and micro-level (individual physical attributes) predictors of Olympic success. GDP gives us an idea for support and training resources, while looking at height 'non-physcially demanding sports' looks at it certain human factors affect sports where raw strength is less relevant.

---

2. Import Packages & Cleaned CSV file

In [17]:
```
import numpy as np
import requests, pandas as pd, time
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression, LogisticRegression
import statsmodels.api as sm
import duckdb
import requests
from bs4 import BeautifulSoup
import time
from scipy.stats import ttest_ind
```

In [18]:
```
merged = pd.read_csv('Cleaned_df.csv')
```

In [19]:
```
country_year = pd.read_csv('country_year.csv')
```

---

3. Data Collection & Cleaning

Cleaning Summary:
During cleaning, we transformationed our data for analysis. We started by filtering both datasets to the years 1992–2016 so they lined up and had consistent coverage. Then we cleaned up the GDP column names by stripping whitespace and reshaped the GDP data from wide to long format so it could be merged properly. We also converted the Year column to integers and standardized country names by lowercasing and removing extra spaces to make sure merges worked smoothly. For missing data, we used a left join so all athlete records stayed in the dataset even if GDP was missing. In specific cases, like when analyzing height or GDP, we dropped rows missing those values but kept the full merged dataset for flexibility later. We also added a numeric "Performance" variable (Gold = 3, Silver = 2, Bronze = 1, None = 0) to make performance comparisons easier.

Our final result is a dataset with ~60,000 athlete event observations from seven Olympic Games, combined with country level economic data.

1.

Load in `GDP.csv` and `athlete_events.csv` with `pandas`. Name the dataframe `GDP_df` and `athlete_events_df`. These csv files have the most of the data we need to make a concrete anaylsis for our research questions. However we will be calling from an API in later steps to finalize our dataset.

Goal: Combine economic and performance data into a single dataset suitable for regression and correlation analysis.

Reasoning: We need both athlete-level and nation-level data to test whether wealth correlates with success and whether height predicts performance within specific sports.

We're working with two main datasets:

- athlete_events.csv: Which contains athlete results from the Olympics, giving data on physical attributes, demographics, and performances.
- GDP.csv: Whcih contains annual GDP data by country in wide format, with each year as a column.

These two dataframes let us to look individual athletic performance with national economic indicators, and look at how wealth correlates with Olympic success.

2.

We mustlimit the data for each set from 1992 to 2016 to best merge both datasets. Additonally some of the column names for GDP had some trailing spaces so we must clean that up as well.

Specific, we clean this data because:

1. The availability of data, since our GDP dataset has complete coverage for these years
2. It's the post-Cold War era, After 1992, we have more stable country definitions (USSR dissolution, Yugoslavia breakup mostly resolved)
3. Things are becoming more modern for the Olympics, and training methods, technology, and professionalization became more standardized.
4. The quality of data is an issue, since physical measurements are more consistently recorded in recent Olympics

3.

Using `SQL` we are going to merge both dataset. Specifically we are going to create a merged df that has all the columns of athlete_df and a new column

`GDP`. We are going to correctly merge GDP on to athlete_event to the correct 'Team' (Mergeded by Country). Before we go ahead and do the merge we must make some changes to `filtered_GDP`.

- In `filtered_GDP` each year is a separate column. but `filtered_athelete` we only have a single `Year` so we must unpivot the wide GDP table into a long, tidy format which we can do with `melt()`
- Additionally, after we melt I noticed that `filtered_GDP`'s Year column is made up of strings which to make the merge work and use for our anaylsis `Year` should be made up of integers. So we change the type to integer.

Goal: Attach GDP values to each athlete record by matching country and year.

Why: This join allows use to look at the GDP data with athlete performance. We use LEFT JOIN to make sure that every athlete in the dataset, even if GDP data is missing for their country or year.

Reshaping GDP Data

Our GDP data has this setup:

Country	1992	1993	1994	...
France	1234	1245	1267	...

We need to change the format to match with athlete data:

Country	Year	GDP
France	1992	1234
France	1993	1245

This allows us to join on country and year, which lets us give each athlete the GDP value from their specific Olympic year.

The merged dataset now has one row for each athete, with their country's GDP for that year.

From our print out, we can begin to notice that athletes from France and Italy have relatively high GDP values (~37,000-40,000) and athletes from Ethiopia have much lower GDP values (~1,250). We used LEFT JOIN to keep all athletes, even if GDP data is missing for their country, which lets us look at how resources correlate with the athletic succes of their athletes.

4.

The function `fetchwb_indicator` is a helper function that fetch data from the World Bank's open API. The World Bank stores global indicators (such as GDP, population, life expectancy, etc.). The function parameters are as follows:

- `indicator` -> The World Bank indicator code you want.
- `country` → ISO code or "all' for every country (default = all).
- `start` and `end` → year range to fetch.

The output of the function is: A pandas DataFrame containing country, ISO code, year, and value for that indicator. While we have total GDP from our CSV, the World Bank provides additional indicators that give More context. It lets use see GDP per capita, which is us adjusting for population size, making the comparisons between small and large countries fairer. It also lets us calculate the per-capita Olympic success rates, and is a very trustworther dataset.

These variables are a good additions as they help us understand if the Olympic success comes from total resources- large countries with high total GDP- or per-capita wealth as a smaller, wealthy country.

Using our function `fetchwb_indicator` we use to create dataframes for the indicators 'Total Population GDP'(df called `pop_df`) and 'GDP per Capita' (df called `gdp_pc_df`) within the respective time frame 1992 to 2016.

Both are useful in different ways:

- `pop_df` measures a country's overall economic output. It might make sense that countries with high total GDP can afford more athletes, training facilities, and sports programs due to their big size and capita.
- `gdp_pc_df` looks at the average wealth per person in a country. This is important to consider because high per-capita GDP countries people might have better access to nutrition, healthcare, and training for the average citizen who might become an athlete.

We think that both might matter, but in different ways. Total GDP might predict number of medals, with more overall resources leading to more athletes, while GDP per capita might predict individual athlete performance, since they might have better training quality.

We then make a single dataframe `wb_df` by performing an `OUTER` merge on `pop_df` and `gdp_pc_df`. This is too make our final merge easier.

Different data sources spell, punctuate, or capitalize country names slightly differently and pandas' merge is case-sensitive and exact. So we normalize both columns before merging this step makes sure they're in the same consistent format.

We do this to account for difference in things like capitalization, where we want "FRANCE" vs "France" vs "france" to all be counted the same, Spacing, so that "United States" and "United States " (trailing space) are both te same, and spelling differents, so that we look at "Korea," "South Korea," and "Republic of Korea."

By converting all names to lowercase and stripping whitespace, we ensure that "France " and "france" are treated as the same country, which helps ensures data is merges correctly.

5.

Finally, we must join your main dataset `merged` with out final World Bank dataset `wb_df` which contains population and GDP per capita by country and year. In doing so we add new columns, Population and GDP_per_capita, to each athlete row, matching on country and year. We do this with `LEFT JOIN` with pandas. This results in our final dataset that would be used for our analysis. Which is `merged` dataframe which we call from the Cleaned_df csv file.

Cleaning Explanation:

- Renamed the columns so that there is consistency across datasets
- Converted `Year` data to integers for merging with GDP data
- Dropped rows with missing `Height` for analyzing individual performance to avoid skew

6. Additionally data cleaning for Hypothesis 1

After merging in GDP data, we further cleaned the dataset to prepare it for analysis. First, we filtered the data to include only medal winning observations, since our outcome is based on medal counts. We then created an indicator for whether each medal came from a less physically demanding sport (e.g., archery or shooting) and aggregated the data to the country year level to compute total medals and the proportion earned in these less-physical sports. Finally, we merged in GDP per capita for each country year and grouped countries into Low, Medium, and High GDP categories using tertiles. Named the dataset

`country_year` and made is csv file. This cleaned and aggregated dataset was used for our regression analysis.

---

4. Data Description

1. What are the observations (rows) and the attributes (columns)?

Each observation represents one athlete's participation in a specific Olympic event in their respective year and with that athlete's country's GDP for that year, GDP per capitia for that year and population of that year. As for the attributes, each attribute is a variable describing the athlete attributes(such as weight, height, sex, etc), event information (Event, Year, Season, Sport, etc), their performance on the Sport(Medal won or NaN) or the athelete's country's GDP, GDP per capitia and population.

2. Why was this dataset created?

The reason behind creating this dataset was to investigate how socioeconomic and physical factors influence athletic performance across different Olympic sports. Specifically, we wanted to explore whether success in sports that rely heavily on technique and precision, such as archery, follows the same patterns as sports that are typically associated with physical dominance, such as basketball. In essence, the dataset was created to bridge the gap between athletic performance data and economic indicators, allowing us to uncover patterns that might reveal how access to resources, training infrastructure, and national wealth shape global athletic success. By focusing on both physically demanding, resourceintensive sport and Archery a precision-based, technique-heavy sport, we can highlight whether economic strength translates differently across contrasting athletic domains.

3. Who funded the creation of the dataset?

Both datasets used in this project were made publicly available on Kaggle by independent contributors and were not part of any formally funded research initiative. Kaggle hosts open, community-shared datasets contributed by researchers, data enthusiasts, and organizations for educational and analytical purposes. 'Athelete_events_df' came from

https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results Which it's contributor is, Heesoo Kim. The data itself originates from public historical Olympic records. 'GDP_df' came from https://www.kaggle.com/datasets/nitishabharathi/gdp-per-capita-all-countries Which it's contributor is Nitisha Bharathi. The dataset compiles publicly available economic data. The `wb_df` dataset was generated by directly retrieving data from the World Bank Open Data API, which provides free and publicly accessible global development indicators. The World Bank, an international financial institution funded by contributions from its member countries, collects and maintains a wide range of economic, social, and demographic data through its various research and development programs.

4. What processes might have influenced what data was observed and recorded and what was not?

There are a lot of things we need to consider that might have influenced the data. Olympic qualification standards might have, since only athletes who met specific qualification criteria and actually competed are included in the records. Athletes who qualified and then later withdrew, or those who competed in demonstration events, could be missed. Another thing might be physical measurements (like height, weight) are more complete for recent Olympics. Earlier games have more missing data due to less systematic data collection practices and evolving privacy standards. Another could be political or geographic changes. We know that country representation changed significantly during the times out data came from (1992-2016) due to a few factors such as: - Dissolution of nations (USSR, Yugoslavia, Czechoslovakia) - Formation of new countries - Changes in NOC (National Olympic Committee) recognition This could have effected both athlete records and GDP data. GDP data collection wcould have also influenced directly since economic data quality varies based on factors like national statistical capacity, honestly, system differences (like a market vs. transitional economies), and things like wars or economic events.

5. What preprocessing was done, and how did the data come to be in the form that you are using?

First we filtered the athlete_events data to only include years from 1992-2016 since that's where we had GDP data. We also grabbed just those year columns from the GDP dataset. Then we cleaned up the GDP dataset a bit since there were some trailing spaces in the col names that we removed using .str.strip(). The GDP data was in wide format with each year as its own column, so we hat to use

pd.melt() to reshape it into long format with separate Year and GDP columns. We also converted the Year column to integers so it would merge properly. To combine the datasets, we used DuckDB to do a LEFT JOIN and matched with both the Team and Country name and the Year. We used LEFT JOIN so we'd keep all the athlete records even if some countries didn't have GDP data(This became our `merged` dataframe. For our analysis, we created a Performance variable that scores medals numerically, with Gold=3, Silver=2, Bronze=1, and no medal=0. This made it easier to compare performance across athletes. Then, when we did specific analyses, we dropped rows with missing values. For the height analysis we dropped rows missing height data, and for GDP analysis we dropped rows missing GDP. But we kept the full merged dataset for other stuff we might want to look at. Additionally we decideded to expand our dataset with columns of GDP per capita and population. To do this we created a function that would call the World Bank API and fetch the data/indicators that we want to add(GDP per capitia and Population). This created a dataframe for each indicator, so we performed an OUTER merge on both of them to get one dataframe. For our final merge we had to nomalize our `merged` and new dataframe representation of country/team to make sure that they have the same coherent format to make the final merge easier. Finally we performed a LEFT JOIN merge to add this new additionall columns onto our `merged` dataframe.

6. If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

The people involved are Olympians who know their results are publicly released and often used for analysis.

7. Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted on Github, in a Cornell Google Drive or Cornell Box).

‘Athelete_events_df’ came from https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results Which it’s contributor is, Heesoo Kim. The data itself originates from public historical Olympic records.

‘GDP_df’ came from https://www.kaggle.com/datasets/nitishabharathi/gdp-per-capita-all-countries Which it’s contributor is Nitisha Bharathi. The dataset compiles publicly available economic data.

'wb_df' dataset was created by directly calling the World Bank Open Data API.

Our GitHub: https://github.com/RachelBethke/final_project_2950 Contains the

raw CSV files, jupyter notebooks, analysis code, and documentation in the main branch.

All of the data is publicly available.

---

5. Data Analysis

Analysis 1: Does Height Predict Success in Archery?

The correspondence of height of and athlete and performance in archery

Purpose: Archery is a precision and technique-based sport where raw physical attributes like height or strength are thought to be less important than in sports like basketball or weightlifting.

Method: We filter the dataset to archery athletes with height data, score performance numerically (Gold=3, Silver=2, Bronze=1, No medal=0), and examine the relationship between height and average performance.

We want to look at if attributes like height still predict success in archery, even though it's not a physically dominant sport.

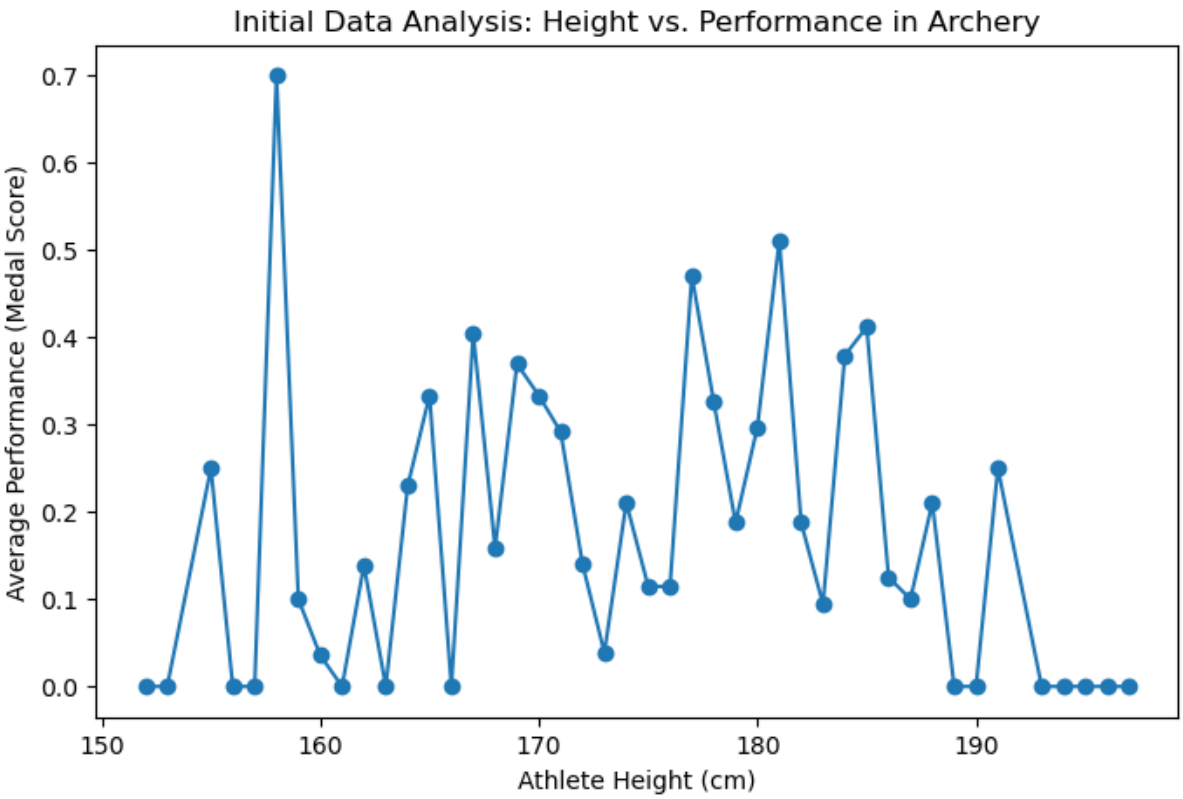
Speifically, we're looking for:

- If height has no effect, we'd see a flat line (performance doesn't change with height)
- If height does matter, we might see taller or shorter archers performing better on average

This will help us understand whether physical attributes universally predict Olympic success or only in specific types of sports.

In [20]:
```
archery_df = merged[merged['Sport'] == 'Archery'].copy()

archery_df = archery_df.dropna(subset=['Height'])

medal_map = {'Gold': 3, 'Silver': 2, 'Bronze': 1}
archery_df['Performance'] = archery_df['Medal'].map(medal_map).fillna(

# Group by exact height value and calculate mean performance for each
# This aggregation shows whether specific heights correlate with bette
height_perf = (
    archery_df.groupby('Height')['Performance']
```

```
    .mean() # Average performance score at a height
    .reset_index() # Convert back to DataFrame format so it can be plo
    .sort_values('Height') # Order from shortest to tallest
)

plt.figure(figsize=(8, 5))
plt.plot(height_perf['Height'], height_perf['Performance'], marker='o'
plt.title('Initial Data Analysis: Height vs. Performance in Archery')
plt.xlabel('Athlete Height (cm)')
plt.ylabel('Average Performance (Medal Score)')
plt.show()
```



Key Finding: We can see that the relationship between height and archery performance is highly variable with no clear linear trend (150-200cm).

Observations:

- There's a notable spike around 158cm, but this is likely due to small sample size at that height
- Most of the plot hovers around 0.1-0.4 average performance across different heights
- There's no obvious pattern suggesting taller or shorter archers consistently perform better

Interpretation: Height may not be a strong predictor of success in archery. However, the high variance suggests we should be cautious—small sample sizes at specific heights create noise. In our final analysis, we might need to group

heights into ranges, look at sample sizes for each height bin, or consider using regression with confidence intervals instead of averaging by exact height to make sure data is usable for analysis.

Analysis 2: GDP Impact Across Sport Types

Having examined height in a precision sport, we now compare how GDP relates to performance across different sport types.

Purpose: Compare how economic resources (GDP) correlate with performance in physically demanding sports (Basketball) versus precision-based sports (Archery).

Method: We aggregate performance by year, sport, and country, then plot average GDP against average performance for both sports to identify patterns.

These sports represent two extremes: Basketball is Physically demanding, requires height, athleticism, specific courts, and seems to favor wealthy countries.

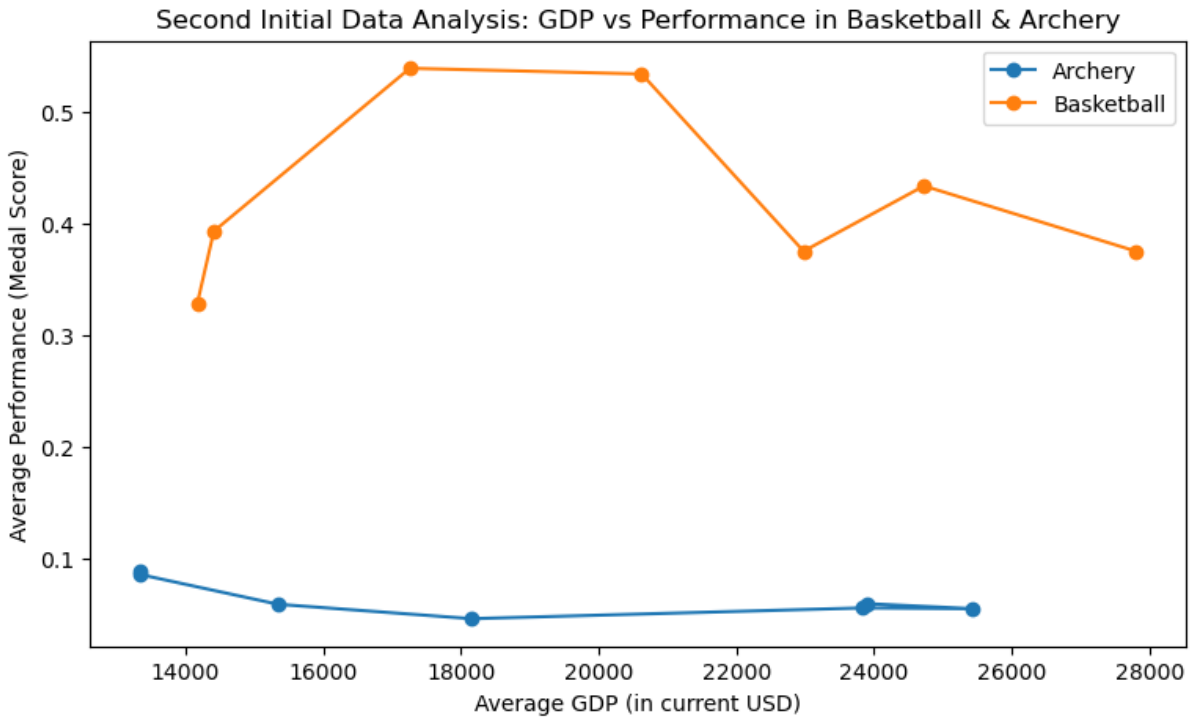
Archery is very technique-based, with less of an obvious dependance on any specific physical traits. It also requires less expensive training facilities, and seems to historically have some success from countries across income levels.

We can begin to form a hypothesis that maybe if GDP matters more for resource-intensive sports, we should see basketball as having a stronger positive relationship between GDP and performance, and archery having a weaker relationship between GDP and performance

Our comparison shows how economic resources translate differently across sport types.

In [21]:
```
sports_df = merged[merged['Sport'].isin(['Archery', 'Basketball'])].co

medal_map = {'Gold': 3, 'Silver': 2, 'Bronze': 1}
sports_df['Performance'] = sports_df['Medal'].map(medal_map).fillna(0)

sports_df = sports_df.dropna(subset=['GDP'])

# Get country-level averages by year and sport
perf_gdp = (
    sports_df.groupby(['Year', 'Sport', 'Team'], as_index=False)
    .agg({'GDP': 'mean', 'Performance': 'mean'}) # Average GDP and per
) # Combined records from same country/year/sport into single data poi

# Get yearly averages across all countries for each sport
```

```
avg_perf_gdp = (
    perf_gdp.groupby(['Year', 'Sport'], as_index=False)
    .agg({'GDP': 'mean', 'Performance': 'mean'}) # Average for all cou
) # Shows overall patterns of sports at a level instead of individual

plt.figure(figsize=(9, 5))
for sport in ['Archery', 'Basketball']:
    subset = avg_perf_gdp[avg_perf_gdp['Sport'] == sport]
    plt.plot(subset['GDP'], subset['Performance'], marker='o', label=s

plt.title('Second Initial Data Analysis: GDP vs Performance in Basketb
plt.xlabel('Average GDP (in current USD)')
plt.ylabel('Average Performance (Medal Score)')
plt.legend()
plt.show()
```



Key Findings:

Basketball (Physically Demanding Sport):

- Performance peaks at mid-range GDP (~$18,000-20,000)
- Higher performance overall compared to archery across most GDP levels
- Suggests GDP matters, but the relationship isn't simply "more money = more medals"

Archery (Precision-Based Sport):

- Performance remains relatively flat across all GDP levels (~0.05-0.10)
- Little variation with GDP changes
- Suggests GDP may be less important for success in technique-based sports

Analysis:
Interpretation: Mid-GDP countries may have optimal combinations of resources and Olympic prioritization for basketball. Very wealthy nations might allocate fewer relative resources to Olympic sports. For archery, low equipment and training costs allow countries at any economic level to compete effectively.
Implication: The type of sport matters significantly—GDP predicts success more strongly in resource-intensive sports than in precision sports.

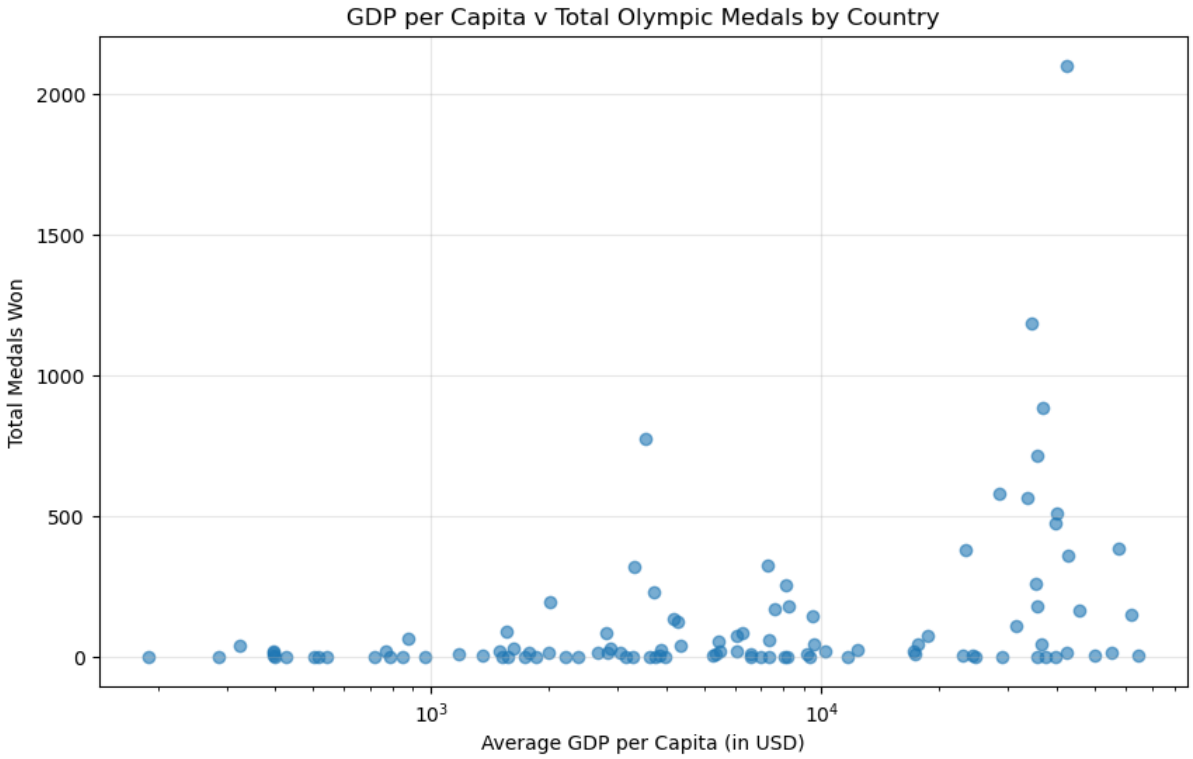
Analysis 3: Country-Level Wealth and Medal Success

Purpose: Examine whether wealthier countries (measured by GDP per capita) win more Olympic medals overall.

Method: We aggregate total medals by country and calculate average GDP per capita, then visualize this relationship using a log scale to handle the wide range of GDP values (from ~$400 to 50,000+).

Why log scale: GDP per capita spans two orders of magnitude across countries, so a logarithmic scale makes patterns more visible.

In [22]:
```
# Aggregate medals by country
country_medals = (merged[merged['Medal'].notna()]
    .groupby('Team')
    .agg({ 'Medal':'count','GDP_per_capita':'mean'}) # Total medals wo
    .rename(columns={'Medal':'Total_Medals'}).dropna() # Removes count
) # Groupd by country to get the total medals and average GDP per capi

plt.figure(figsize=(10,6))
plt.scatter(country_medals['GDP_per_capita'],country_medals['Total_Med
plt.xlabel('Average GDP per Capita (in USD)')
plt.ylabel('Total Medals Won')
plt.title('GDP per Capita v Total Olympic Medals by Country')
plt.xscale('log')
plt.grid(True, alpha=0.3)
plt.show()
```



Key Findings:

- Positive correlation: Countries with higher GDP per capita tend to win more medals overall.
- High Variance: Countries with similar GDP per capita show dramatically different medal counts.
- Wealth doesn't guarantee success: Some small wealthy nations have high GDP per capita but few medals.
- Outliers exist: A few countries (USA, Russia, China, Germany) dominate with 500-2,000+ medals.

We can consider that wealth alone doesn't promise Olympic success. Other factors like **population size**, **sports culture**, and **historical Olympic importance for a country** can have a large effect on success. Some smaller wealthy nations have high GDP per capita but few medals, while larger nations with moderate GDP per capita do well maybe just due to the total number of athletes.

Analysis 4: Medal Distribution by Economic Tier

Purpose: Determine if high-GDP countries win proportionally more medals than low-GDP countries.

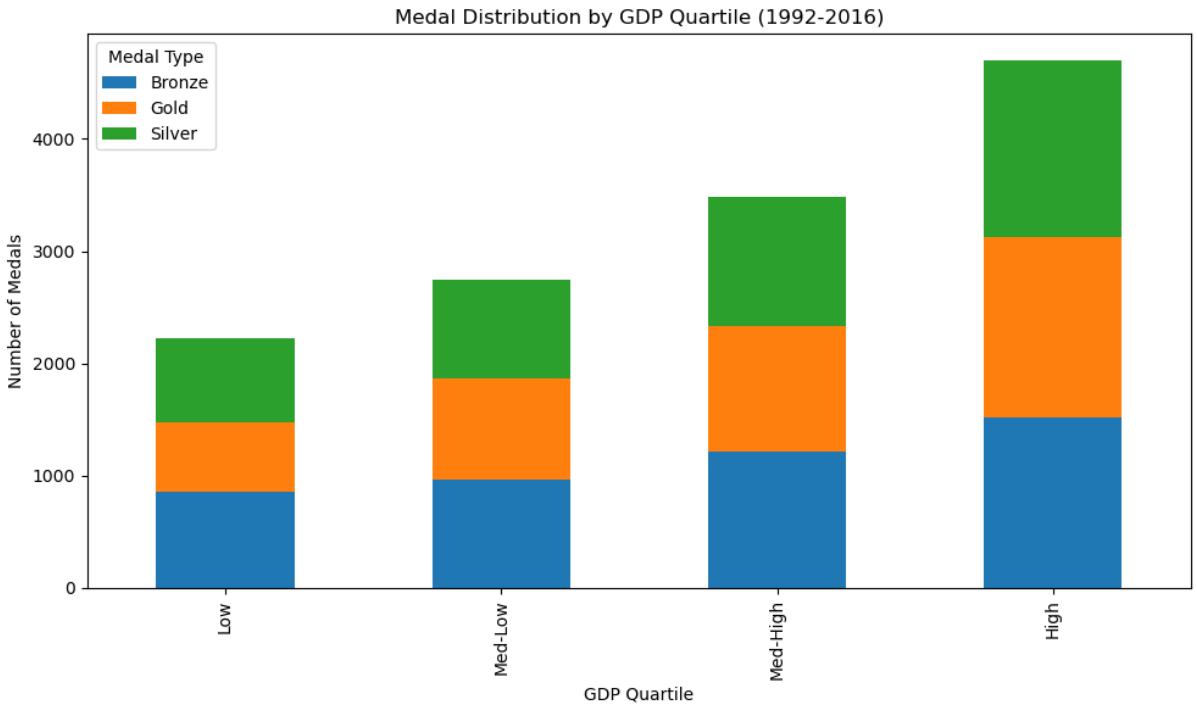
Method: We divide countries into four equal groups (quartiles) based on GDP:

- Low: Bottom 25% of GDP values

- Medium-Low: 25th-50th percentile
- Medium-High: 50th-75th percentile
- High: Top 25% of GDP values

Then we count medals won by each group and visualize the distribution. Grouping reduces noise from year-to-year GDP fluctuations and creates balanced comparison groups for clearer pattern detection.

In [23]:
```
# Count medals with GDP quarters
merged_with_gdp = merged.dropna(subset=['GDP']).copy()
merged_with_gdp['GDP_Quartile'] = pd.qcut(merged_with_gdp['GDP'],
                                          q=4,  #splits data into qu
                                          labels=['Low', 'Med-Low',
medal_counts = (
    merged_with_gdp[merged_with_gdp['Medal'].notna()]
        .groupby(['GDP_Quartile', 'Medal'], observed=False) # Ensures
        .size() # Counts the number of medals in each quartile-medal c
        .unstack(fill_value=0) # Pivot us to a wide format
)

medal_counts.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Medal Distribution by GDP Quartile (1992-2016)')
plt.xlabel('GDP Quartile')
plt.ylabel('Number of Medals')
plt.legend(title='Medal Type')
plt.tight_layout()
plt.show()
```



The graph shows higher GDP correlating with more total medals.

Key Findings:

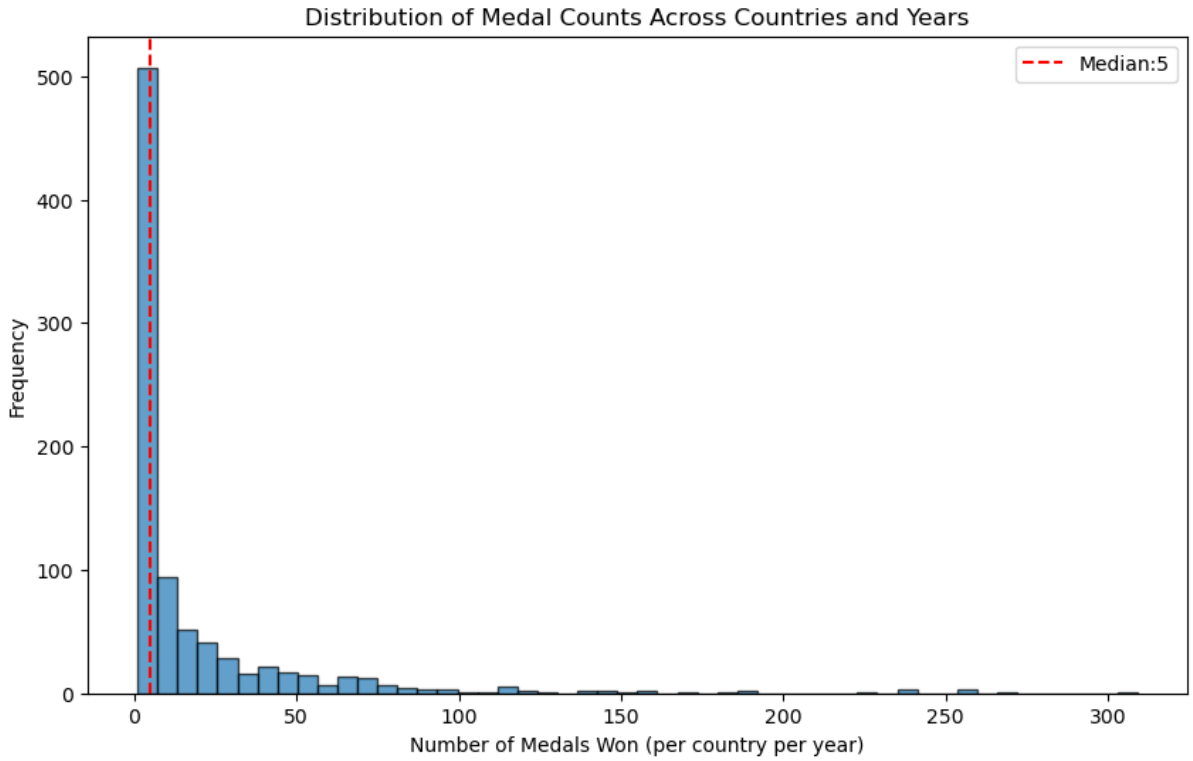
- High GDP quartile has won ~5,000 total medals.
- Low GDP quartile has won ~2,200 medals, which is less than half.
- The gradient is very consistent across all medal types (Bronze, Silver, Gold).
- Each step up in GDP quartile shows an almost consistant increase in medals won.

Important Considerations:

- This doesn't prove that GDP causes success.
- Confounding factors include population size, culture, and historical Olympic participation.
- To isolate GDP's effect, we'd need to analyze medals per capita or per athlete sent.

We can also make sure to look at averages over all countries to get a baseline.

In [24]:
```
# Look at how many medals countries win on average
medals_per_country = (
    merged[merged['Medal'].notna()]
    .groupby(['Team','Year']).size() # Count medals per country per ea
    .reset_index(name='Medal_Count')
) # This shows the distribution of success, so we can see if most coun

plt.figure(figsize=(10,6))
plt.hist(medals_per_country['Medal_Count'], bins=50, edgecolor='black'
plt.xlabel('Number of Medals Won (per country per year)')
plt.ylabel('Frequency')
plt.title('Distribution of Medal Counts Across Countries and Years')
plt.axvline(medals_per_country['Medal_Count'].median(),color='red',lin
plt.legend()
plt.show()
```



We can see that most countries win very few medals, while a few dominate. This context is important for understanding inequality in Olympic performance.

This histogram reveals extreme inequality in Olympic success. The distribution of medals is very uneven. The median is 5, meaning half of all country-year results are 5 medals or fewer. Most countries win just a handful, while a small group racks up hundreds. This creates a long tail in the data—those few outlier performances of 300+ medals stretch the distribution far to the right.

We can start to interpret this data as showing that olympic success is highly concentrated. Most participating countries win a handful of medals, while a small amount of countries consistently dominate the medal count.

This is important context to keep in mind during our analysis since, if we find that GDP correlates with medals, we need to think about if this might be because wealthy countries send more athletes, wealthy countries have better training access, and that the same wealthy countries dominate both GDP rankings and Olympic results.

To check which countries are the handful that dominate, we write code to grab that information.

If GDP correlates with medals, it's very important to know what countries these are becayse these same countries are also the world's largest economies, GDP lets a country to boost performance, and the success is driven by factors that

correlate with both GDP and medals.

In [25]:
```
# Prove which countries dominate Olympic medals
print("="*60)
print("TOP 15 COUNTRIES BY TOTAL MEDALS (1992-2016)")
print("="*60)

top_countries = (merged[merged['Medal'].notna()]
    .groupby('Team').size()
    .sort_values(ascending=False).head(15)
    .reset_index(name='Total_Medals')
)

print(top_countries.to_string(index=False))
print("="*60)

total_medals=merged['Medal'].notna().sum()
top_10_medals=top_countries.head(10)['Total_Medals'].sum()
top_10_percentage = (top_10_medals/total_medals)*100

print(f"\nTotal medals awarded (1992-2016):{total_medals}")
print(f"Medals won by top 10 countries:{top_10_medals}")
print(f"Percent of medals won by top 10:{top_10_percentage:.1f}%")
print("="*60)
```

```
============================================================
TOP 15 COUNTRIES BY TOTAL MEDALS (1992-2016)
============================================================
         Team  Total_Medals
United States          2101
      Germany          1185
       Russia          1100
    Australia           885
        China           777
       Canada           716
        Italy           584
       France           566
Great Britain           556
  Netherlands           514
        Japan           479
  South Korea           447
       Norway           387
        Spain           380
       Sweden           362
============================================================

Total medals awarded (1992-2016):16781
Medals won by top 10 countries:8984
Percent of medals won by top 10:53.5%
============================================================
```

We can now see that the small amount of countries that consistently dominate

the medal count are the USA, Germany, Russia, Australia, China, Canada, Italy, France, Great Britain, and the Netherlands.

Having explored GDP and height patterns across sport types, we now move to our preregistered hypotheses to test these relationships formally with statistical rigor.

Our exploratory analyses revealed interesting patterns: GDP correlates with total medals, but the relationship varies by sport type. Height shows different impacts across sport categories. These observations motivated our two preregistered hypotheses, which we now test formally with appropriate statistical methods.

---

6. Preregistration Statements

Hypothesis 1:

Countries with a lower GDP perform better (obtain more total medals) in less Physical intense sports(e.g Rifle shooting, archery) then higher GDP countries.

Description of Analysis:

Run a linear regression where we input GDP group (as a dummy variable: Low GDP, Medium GDP, High GDP) and output the proportion of medals earned in determined less physically demanding sports for each country in a given Olympic Games. The High GDP group will serve as the reference category, so we will test whether $\beta_{lowGDP} > 0$, indicating that lower-GDP countries earn a higher proportion of their medals from non-physical sports compared to higher-GDP countries.

Context: Economic development has long been discussed as a potential driver of Olympic performance. Wealthier countries typically have greater access to training facilities, national sports programs, advanced coaching, and sports science, which can lead to stronger performance in physically demanding events such as track, swimming, and team sports. In contrast, less financially intensive sports such as archery, shooting, or other precision-based events—require smaller team sizes, less specialized infrastructure, and lower long term investment. Hence, we belive that lower GDP countries may earn a larger share of their medals in less physically demanding sports, while higher GDP nations may dominate in strength and endurance based events that require substantial

institutional support.

Data Analysis 1:

In [26]:
```
y = country_year["prop_less_physical"].astype(float)

# Create dummy variables for GDP groups (Low, Medium, High)
GDP_dummies = pd.get_dummies(
    country_year["GDP_group"],
    prefix="GDP",
    drop_first=True # Makes 'High' GDP the reference category, omittin
) # Coefficients show difference from High GDP group

# Adds constant term, or the intercept, required for OLS regression
X = sm.add_constant(GDP_dummies.astype(float))
ols_model = sm.OLS(y, X).fit()

print(ols_model.summary())
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:     prop_less_physical   R-squared:
0.006
Model:                            OLS   Adj. R-squared:
0.003
Method:                 Least Squares   F-statistic:
1.764
Date:                Fri, 05 Dec 2025   Prob (F-statistic):
0.172
Time:                        17:58:44   Log-Likelihood:
426.82
No. Observations:                 577   AIC:
-847.6
Df Residuals:                     574   BIC:
-834.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
=======
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
const          0.0342      0.008      4.098      0.000       0.018
0.051
GDP_Low        0.0042      0.012      0.356      0.722      -0.019
0.027
GDP_Medium    -0.0167      0.012     -1.418      0.157      -0.040
0.006
==============================================================================
=======
Omnibus:                      707.513   Durbin-Watson:
1.720
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            50
213.662
Skew:                           6.185   Prob(JB):
0.00
Kurtosis:                      46.995   Cond. No.
3.73
==============================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

In [27]:
```
params = ols_model.params
conf_int = ols_model.conf_int()
pvalues = ols_model.pvalues

beta_low = params.get("GDP_Low", np.nan)
```

```
ci_low = conf_int.loc["GDP_Low"] if "GDP_Low" in conf_int.index else (
p_low = pvalues.get("GDP_Low", np.nan)

print("β_lowGDP:", beta_low)
print("95% CI for β_lowGDP:", tuple(ci_low))
print("p-value for β_lowGDP:", p_low)
```

```
β_lowGDP: 0.004202503614957722
95% CI for β_lowGDP: (-0.019007085929482692, 0.027412093159398132)
p-value for β_lowGDP: 0.7222440083953661
```

Evaluation of significance:

To evaluate our preregistered hypothesis, we tested whether low-GDP countries earn a higher proportion of their medals in less physically demanding sports than high-GDP countries by examining the coefficient β_lowGDP in our regression model. The estimated coefficient was β_lowGDP = 0.004 with a 95% confidence interval of (–0.019, 0.027) and a p-value of 0.72. Because the confidence interval includes zero and the p-value is large, the observed data provide no statistically significant evidence that low-GDP countries earn a larger share of their medals from less-physical sports. Therefore, we fail to reject the preregistered hypothesis's null, concluding that within our dataset, the proportion of medals coming from less-physical sports does not differ meaningfully between low-GDP and high-GDP countries.

Conclusion for Hypothesis 1

Research Question: Do lower-GDP countries earn a higher proportion of medals in less physically demanding sports (archery, shooting) compared to higher-GDP countries?

Answer: No.

Our research question asked whether a country's GDP correlates with its performance in the Olympic Games, and which types of sports tended to favor lower-GDP versus higher-GDP countries. To address this, we examined whether countries with lower GDPs earned a larger share of their total medals in less physically demanding sports such as shooting and archery. Our regression results showed that the difference between low-GDP and high-GDP countries was extremely small (β_lowGDP = 0.004) and not statistically significant (p = 0.72), with a confidence interval centered near zero. This indicates that lower-GDP countries do not systematically win a larger proportion of medals in non-physical sports compared to higher-GDP nations.

Overall, based on the specific analysis conducted, we find no meaningful

evidence that GDP is related to specialization in less-physical Olympic sports, nor that low-GDP countries perform relatively better in these sports. While GDP may still relate to overall medal counts or performance in other sport categories, our analysis suggests that differences in economic development do not translate into advantages in less physically demanding Olympic events.

Hypothesis 2:

In more physically demanding sports, better performing athletes (win a medal) will be taller, on average, than shorter athletes, whereas in less physically demanding sports, there will be no meaningful correlation (close to zero).

Description of Analysis:

To test this hypothesis, we use the merged dataframe (our cleaned analysis-ready dataset) and perform the following dataset restrictions:

1. Height values must be present. Rows missing height are excluded, since this analysis directly depends on comparing height distributions.
2. Sports must belong to one of two curated categories:
   - Physically demanding: sports where height plausibly provides biomechanical advantage
   - Less physically demanding: sports primarily determined by technique, stability, or fine motor control.
3. Medal indicator: We convert medal status into a binary variable:
   - 1 = medalist (Gold/Silver/Bronze)
   - 0 = non-medalis
4. Following our preregistration plan, we conduct two independent two-sample t-tests:
   - Medalists vs non-medalists within physical sports
   - Medalists vs non-medalists within non-physical sports

Context: Since medalist and non-medalist sample sizes and variences are both very different, it makes more sense to use Welch's t-test rather than the a normal t-test.

Data Analysis 2:

Physical vs. non-physical catagories are based on the percived biomechanical demands of each sport.

In [28]:
```
height_df = merged.dropna(subset=["Height"]).copy()
```

```
physical_sports = ['Athletics',
   'Basketball',
   'Boxing',
   'Rowing',
   'Swimming',
   'Weightlifting',
   'Wrestling',
   'Handball',
   'Football',
   'Water Polo',
   'Cycling',
   'Volleyball',
   'Hockey',
   'Ice Hockey',
   'Rugby Sevens',
   'Taekwondo',
   'Judo',
   'Tennis',
   'Triathlon',
   'Modern Pentathlon',
   'Synchronized Swimming',
   'Trampolining',
   'Badminton',
   'Softball',
   'Baseball']
non_physical_sports = ['Archery',
    'Shooting',
    'Fencing',
    'Table Tennis',
    'Sailing',
    'Equestrianism',
    'Golf',
    'Curling']
phys_df= height_df[height_df["Sport"].isin(physical_sports)].copy()
nonphys_df= height_df[height_df["Sport"].isin(non_physical_sports)].co

# This converts medal status into format needed for t-test comparison,
for df in [phys_df, nonphys_df]:
    df["is_medalist"] =df["Medal"].notna().astype(int)
```

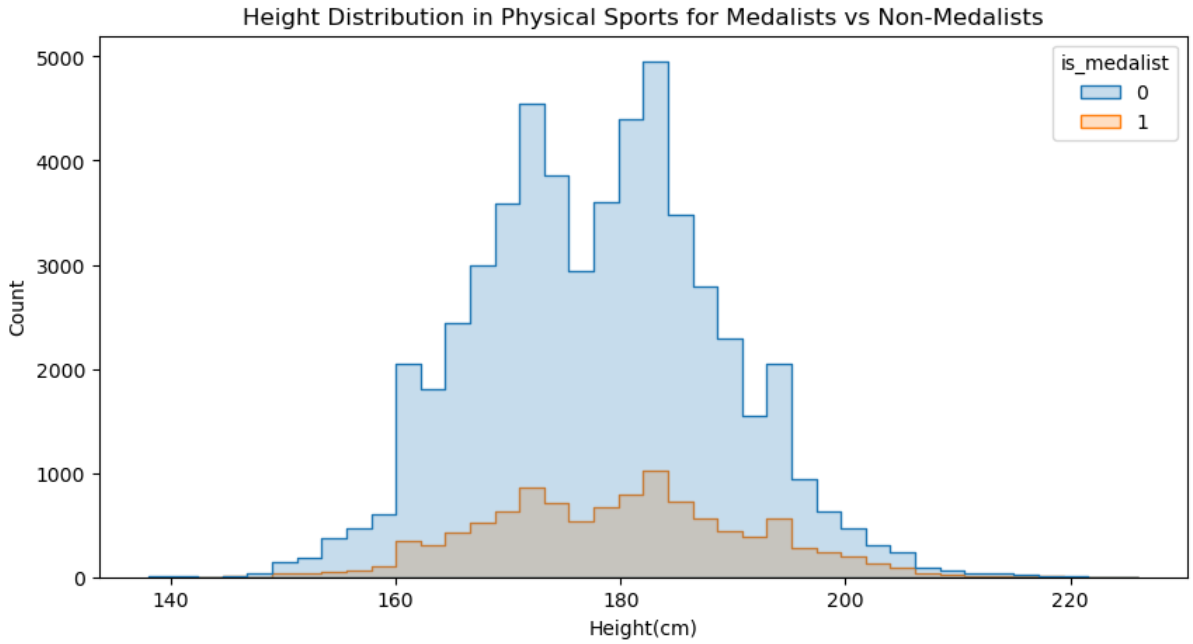
In [29]:

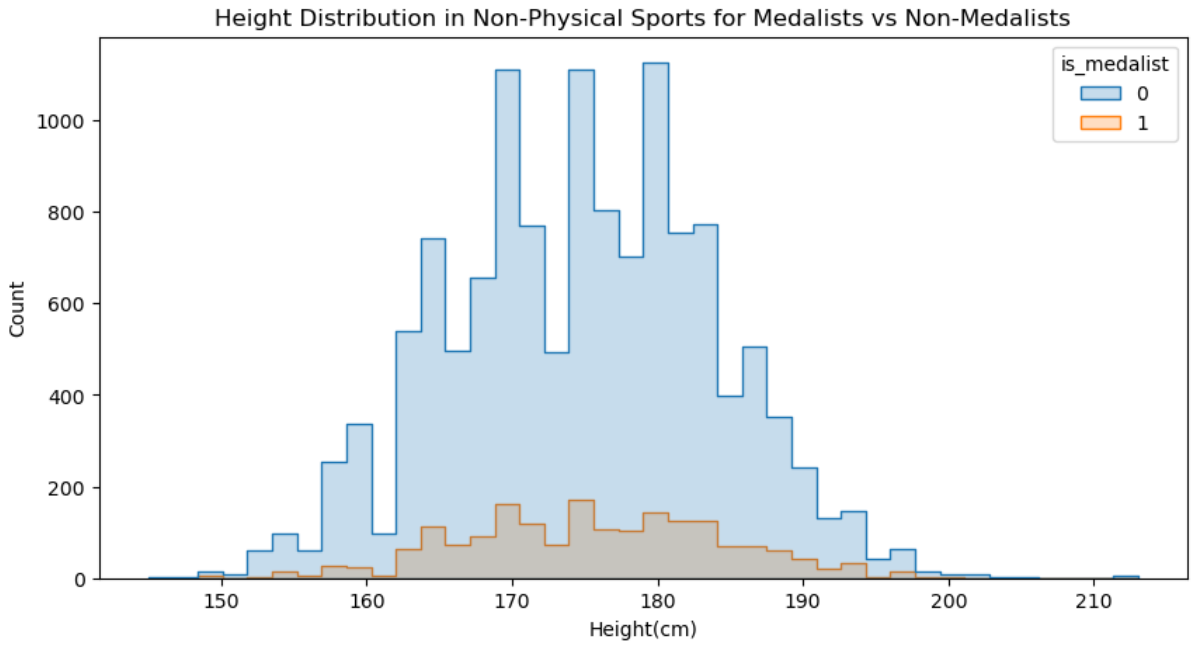
```
def summarize_height_ttest(df, group):
    h_medal= df.loc[df["is_medalist"] == 1,"Height"].astype(float).dro
    h_non= df.loc[df["is_medalist"] == 0,"Height"].astype(float).dropn
    n_medal= len(h_medal)
    n_non= len(h_non)
    mean_medal= h_medal.mean()
    mean_non= h_non.mean()
    std_medal= h_medal.std()
    std_non= h_non.std()

    # Welch t-test
    t_stat, p_val =ttest_ind(h_medal, h_non,
```

```
                                equal_var=False, # Uses Welch's version s
                                nan_policy="omit")

    # Cohen's d, looking at standardized mean difference.
    pooled_var =((n_medal-1) * std_medal**2 +(n_non-1)* std_non**2)
    # Pooled standard deviation allows us to look at both group varian
    pooled_var/=max(n_medal+n_non-2,1) # Avoid division by zero
    if pooled_var>0: pooled_sd = np.sqrt(pooled_var)
    else: pooled_sd=np.nan
    if pooled_sd>0: cohen_d =(mean_medal-mean_non)/pooled_sd
    else: cohen_d=np.nan

    return {"group": group, "n_medal": n_medal, "n_non_medal": n_non,
            "mean_medal_height": mean_medal, "mean_non_medal_height":
            "std_medal_height": std_medal, "std_non_medal_height": std
            "p_value": p_val, "cohen_d": cohen_d }
```

In [30]:
```
phys_results = summarize_height_ttest(phys_df, "Physical sports")
nonphys_results = summarize_height_ttest(nonphys_df, "Non-physical spo
```

In [31]:
```
plt.figure(figsize=(10,5))
sns.histplot(phys_df,x="Height",hue="is_medalist",bins=40,kde=False,el
plt.title("Height Distribution in Physical Sports for Medalists vs Non
plt.xlabel("Height(cm)")
plt.ylabel("Count")
plt.show()

plt.figure(figsize=(10,5))
sns.histplot(nonphys_df,x="Height",hue="is_medalist",bins=40,kde=False
plt.title("Height Distribution in Non-Physical Sports for Medalists vs
plt.xlabel("Height(cm)")
plt.ylabel("Count")
plt.show()
```





In [32]:
```
print("="*60)
print("Physical Sports: Height vs Medal Status")
print("="*60)
for k, v in phys_results.items():
    print(f"{k}: {v}")
print("")

print("="*60)
print("Non-Physical Sports: Height vs Medal Status")
print("="*60)
for k, v in nonphys_results.items():
    print(f"{k}: {v}")
```

```
============================================================
Physical Sports: Height vs Medal Status
============================================================
group: Physical sports
n_medal: 10847
n_non_medal: 54052
mean_medal_height: 179.7029593435973
mean_non_medal_height: 177.8671834529712
std_medal_height: 11.493205005347773
std_non_medal_height: 10.767674897796882
t_stat: 15.33923533536994
p_value: 1.0558525830064431e-52
cohen_d: 0.16853895756892878

============================================================
Non-Physical Sports: Height vs Medal Status
============================================================
group: Non-physical sports
n_medal: 1868
n_non_medal: 12935
mean_medal_height: 175.7093147751606
mean_non_medal_height: 174.78306919211443
std_medal_height: 8.820387043471051
std_non_medal_height: 9.0586516117875
t_stat: 4.228049281147661
p_value: 2.4429707182244245e-05
cohen_d: 0.10258626495381162
```

Summary:

Using the filtered dataset of athletes with non-missing heights, we ran two independent Welch's t-tests comparing medalists and non-medalists within (1) physically demanding sports and (2) less physically demanding sports.

The tests show:

- A statistically significant height difference in physical sports (p~=1e-52)
- A very small but statistically significant height difference in non-physical sports (p~=2.4e-05)

In both groups, the effect size (Cohen's d) is small, but notably larger in physical sports (d ≈ 0.17 vs d ≈ 0.10)

This pattern aligns with our preregistered expectation: height is more predictive of medal performance in physical sports than in non-physical ones

Evaluation of significance:

For **physically demanding sports**, although the effect size is small, the

difference is lanf and directionally consistent with known biomechanical factors. Taller athletes seem to have measurable advantages in many physical Olympic events. This makes sense, since taller athletes have an advantage with stride length in running, leverage in rowing, reach and block height in team sports, hydrodynamics in swimming, etc. The large sample size means even modest differences can reach significance, but the height advantage is real and consistent.

For **non-physically demanding sports**, even though the test reaches statistical significance, the actual effect size seems to be is minimal. A ~1 cm height difference is not meaningfully tied to performance in sports like archery, shooting, table tennis, sailing, or equestrian. These events seem to depend on motor control, precision, steadiness, decision-making, or even animal factors rather than height. The statistical significance of the test more likely comes from the very large sample sizes, small between-group variance. Sinece the Olympics is a pretty speicfic and competitive pool to be getting data from, it makes sense that small differences might sometimes appear across many events, but it doesn't seem as if this indicates an actually meaningful height advantage.

Conclusion 2:

These results match our preregistered expectation: the height difference is larger in physically demanding sports and close to zero in non-physical sports. Both groups show correlations, but effect size is the correct metric, and the physical-sport effect is substantially larger. This makes sense with expectations for sports that rely less on raw power requiring more fine-tuned specific skills.

Combined Conclusion:

Looking at both our hypothesis and analyses, we can see patterns that align with the intuitive relationship between economic resources, physical attributes, and Olympic success. Countries with higher GDP tend to earn a larger share of total medals, while athletes in physically demanding sports show a small but noticeable height advantage among medalists. On the other hand, height seems to play almost no role in the less physical sports. While these trends hold up well in our dataset, they're limited by missing data, simplified performance measures, and economic proxies that can't fully capture training environments. Overall, these findings seen to indicate that both the economic context and physical attributes of teams play roles in Olympic performance, and that this impact is also heavily dependent on if the sport is more physical or technical.

---

7. Data limitations

One limitation of the dataset is that height is missing for about 5% of athletes (6028 of 122216). While this is not a large fraction overall, the missing values are almost definetly not evenly distributed across sports or years. Because we drop rows with missing height, our sample for the t-tests could be biased toward sports or countries that have this data more regularly.

Because our dataset includes Olympics from 1992–2016, these results might just be showing changes over time. The way things are measured, the countries participating, and the amount of people wanting to compete all change accross the decades, so some patterns could be a result of this.

Another limitation is that performance is represented only by whether an athlete received a medal. We do not have information about placements beyond the top three, so athletes who finished fourth or reached finals are treated the same as athletes eliminated early. This is a bit of a simplification of performance since we treat is as a binary outcome. This could be innaccurate or misrepresent the actual relationship between physical attributes and how well athletes actually perform on average.

Additionally, GDP provides only one dimension of a country's economic context and may not fully represent access to sports funding, training infrastructure, or athlete development programs.

---

8. Acknowledgements

We thank the INFO 2950 course staff, especially Evan Vu and Federica Bologna, for their guidance throughout this project.

9. Bibliography

Data Sources:

Kim, H. (2018). *120 years of Olympic history: athletes and results*. Kaggle Dataset. https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-

athletes-and-results

Bharathi, N. (2020). *GDP per capita - all countries*. Kaggle Dataset. https://www.kaggle.com/datasets/nitishabharathi/gdp-per-capita-all-countries

World Bank. (2024). *World Development Indicators*. World Bank Open Data API. https://data.worldbank.org/

Code Repository:

[4] Bethke, R., & Cortez, J. (2024). *Olympic Performance and Economic Indicators Analysis*. GitHub. https://github.com/RachelBethke/final_project_2950
