Introduce the problem -

The problem I am investigating is “Do host nations win significantly more medals when they host the Olympics compared to before and after hosting?”

The questions I want to explore are
Medal Boost:

How many medals does a country win when it is the host compared to the immediately prior and following Olympics?

Is the increase statistically significant?

Gold Medal Effect:

Is the host advantage more pronounced for gold medals vs total medals?

Sport-Specific Advantage:

Do hosts gain more in certain sports (e.g., those requiring large infrastructure like swimming, cycling, gymnastics)?

Introduce the data -

I found the data on kaggle at this link
https://www.kaggle.com/datasets/the-guardian/olympic-games/data?select=dictionary.csv

The data includes
201 unique countries
the respective country code for each country
the population of each country,
the GDP per capita of each country
The year the olympics was held
The host city of the olympics
What sports where participated in at each olympics\
Each athlete that has won a medal
The country the athlete is from
Their gender
The event they participated in
And what medal they earned

Pre processing steps -

I will load the data for dictionary, summer, and winter.csv which includes (medals/athletes/events) and (country codes, population, GDP).
This allows me to understand the structure of the data makes it easier to identify missing values.

The next pre processing step I will take is to handle any missing values

Then I will standardize country codes

Filter out irrelevant columns

Create host country indicator

Add up medal count

Storytelling -

Based on the data from the Olympic games between 1896 and 2012 it appears that there is an overall increase of medals won for the country hosting the games, this boost is present in both the winter and summer games. The most pronounced boosts came from the United States and Norway both of which increased their medal count by over 50% compared to the previous olympic games. Based on the data and graphs there is a significant boost to medals earned when a country hosts the olympic games. There are many possible causes for this one of the first things that comes to my mind is food/diet athletes competing at home are used to the food which prevents them from getting sick or over/under eating which could lead to a boost in performance.

Impact -

The data I collected might be able to encourage countries that normally underperform at the Olympics to host the games in order to improve their medal count and earn more recognition on the world stage. This however does come with some downsides as the Olympic games almost always cost more to host than the host country makes back in tourism revenue/taxes surrounding the games. The draw of an increase in medals might make governments overlook the financial impacts the games can have on a city or country.

References -

The data set I used -
https://www.kaggle.com/datasets/the-guardian/olympic-games?resource=download&select=dictionary.csv

I used this article to help me find the host cities for each olympics
https://en.wikipedia.org/wiki/List_of_Olympic_Games_host_cities

I used ChatGPT to help me aggregate the medal counts

Me: I need help combining the summer and winter medal counts from the olympics
Both the summer and winter csv files have the following headers Year Country Sport Event Medal how would I write code in a jupiter notebook to combine the medals.

ChatGPT:

```
summer_medals = summer.groupby(["Year", "Country").agg(
    Total_Medals=("Medal", "count"),
    Gold_Medals=("Medal", lambda x: (x == "Gold").sum())
).reset_index()

winter_medals = winter.groupby(["Year", "Country"]).agg(
    Total_Medals=("Medal", "count"),
    Gold_Medals=("Medal", lambda x: (x == "Gold").sum())
).reset_index()
```