A Historic Look at the Olympics

By: Erik Hoaglund, Eugene Kikuchi, Priscilla Robinson, and Bob Verner

OLYMPIC FEVER!!

Motivation and Research Questions

With the Summer Olympics coming to a close, it made the topic interesting to look at how the Olympic games have evolved over time.

Research Questions

- How do countries who compete only in summer differ from countries that compete in both seasons of Olympics?
- Does the country's latitude impact their all-time total number of medals won?
- Does the country's latitude impact their all-time average gold medal win percentage?

Hypothesis

Countries farther from the equator will perform better in the Winter Olympics, and countries closer to the equator will perform better in the Summer Olympics.

Questions & Data

Approach

- How do countries who compete only in summer differ from countries that compete in both seasons of Olympics?
  - Found countries that only competed in Summer
  - Found the countries that competed in both the Winter and Summer Olympics
  - Removed data prior to 1924 (First year of Winter Olympics)
  - Focused on Northern Hemisphere countries
- Does the country's location impact total number of medals or gold medals won over time?
  - Find and import dataset that has country's latitude and longitude
  - Create Gold% variable to measure gold medal performance
  - Focused on Northern Hemisphere countries

Datasets

- "120 years of Olympic history: athletes and results"
  - Includes every athlete and event in Olympic history
  - Includes various variables: Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City of Olympics, Sport, Event, and Medal
- "country lat/long"
- country codes

Data Cleanup and Exploration

Exploring and Cleanup

- Additional data exploration - Population
  - Should we really be looking at population instead of latitude?
  - Find a dataset with every country's population and retain
- Merge all datasets
  - athlete_events.csv
  - country_codes.csv
  - country_lat_long.csv
  - noc_regions.csv
  - Population_by_country.csv
- Create dataframe with only countries that have competed in both the Winter and Summer Olympics
- Remove data prior to 1924

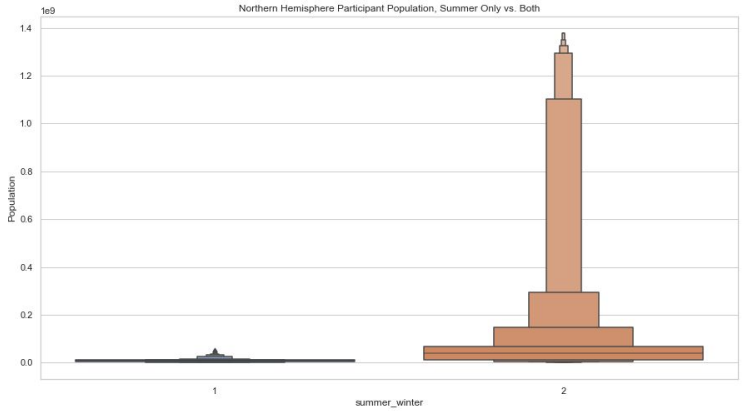
Insights and Problems

- Unable to plot every year due to too many data points; averages and totals suitable for answering our questions

Interesting Developments

Winter and Summer Participation

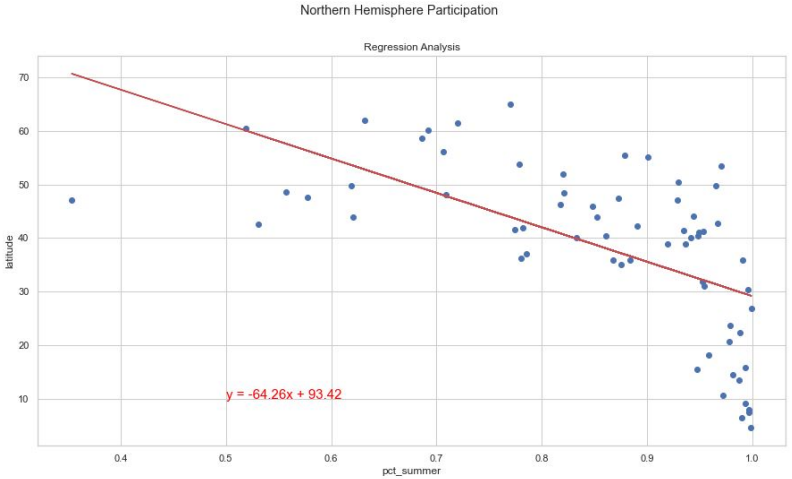
By Population (Seaborn boxenplot)



1. Countries that attend summer only have two characteristics:
    a. They are all close to the equator
    b. Their population is small, with little variance
2. Countries that participate in both have a wide range of populations, and include both small and the largest countries.

Winter and Summer Participation

By Country and Latitude



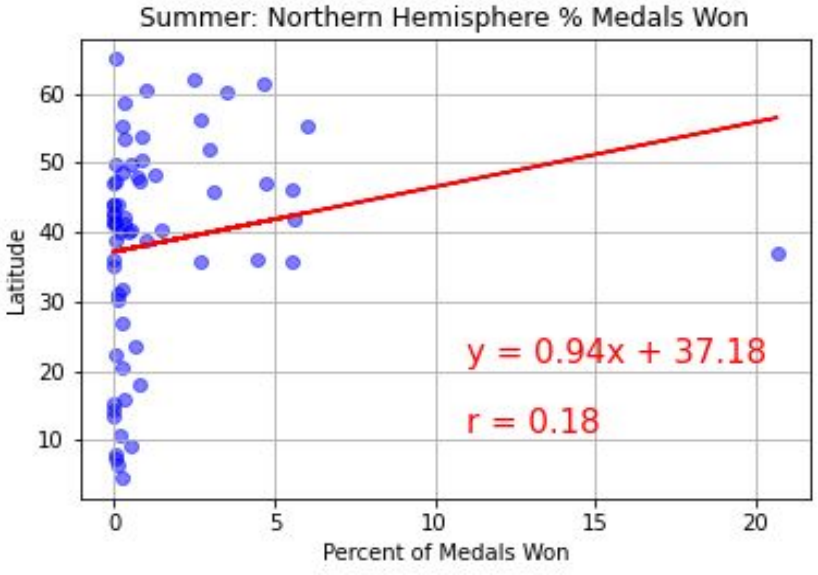
1. For countries below 30 degrees latitude, (i.e. below the southern tip of Florida) participation in the winter olympics drops off considerably
2. In spite of the sharp, non-linear decline in the data points < 30 degrees latitude, the linear regression results are significant ($p<.05$).

Question: Does the Country's Latitude Impact All-Time Total Number of Medals Won?

To analyze country performance by latitude, the data was:

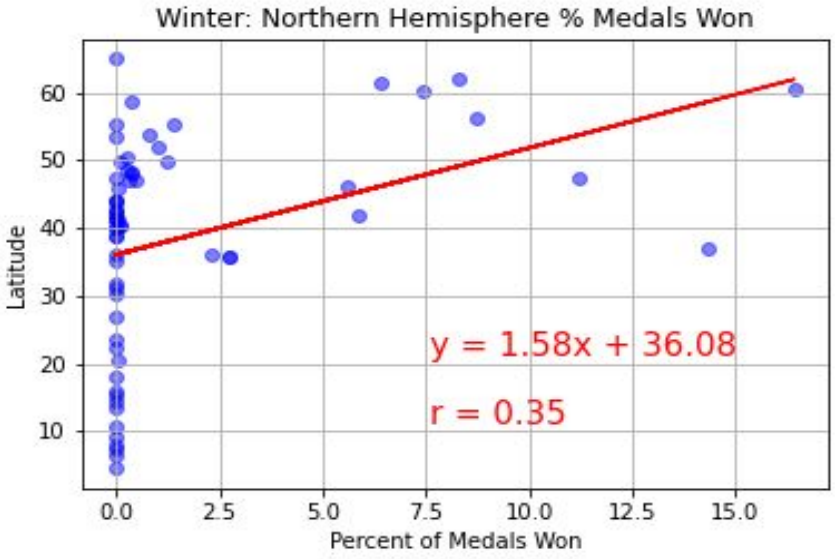
- Separated by winter and summer olympics
- Grouped by Country, Year, and Event
- Aggregated into one medal count for team events
- Summed by total medal count for each country

Olympic Performance by Latitude: Summer



1. With the exception of the outlier, the USA, the majority of medal winning countries are above 30 degrees latitude and have won less than 7% of medals.
2. There is a very weak, positive correlation between latitude and medals won.

Olympic Performance by Latitude: Winter



1. Generally, most of the medals are won by countries greater than 35 degrees latitude (North Carolina ~35)
2. There is only one medal winning country below 35 degrees latitude during the winter olympics.
3. The regression analysis was not significant, likely due to the low win rate in general, and the low to no win rate for lower latitude countries

```
In [ ]:
df_events_f.drop(df_events_f.loc[df_events_f['Season'] == 'Summer'].index, inplace = True)
winter_df = pd.DataFrame(df_events_f)

winter_df
```

```
In [ ]:
wint_gold_df = winter_df.loc[winter_df['gold'] > 0]

wint_gold_df
```

```
In [ ]:
df_events_g = wint_gold_df.groupby(['Year','NOC_x','latitude'], as_index=False).agg({'Name':'count'})
df_events_g = df_events_g.rename(columns={'NOC_x':'NOC'})
df_events_g
```

```
In [ ]:
wint_gold_group = wint_gold_df.groupby(['Year'])

#Create dataframe
wint_gold_group_df = pd.DataFrame()
wint_gold_group_df
wint_gold_group_df['Total Gold'] = wint_gold_group['gold'].sum()

wint_gold_group_df
```

```
In [ ]:
wint_combo_gold = pd.merge(df_events_g, wint_gold_group_df , on=["Year"])

wint_combo_gold
```

```
In [ ]:
wint_combo_gold['Gold %'] = wint_combo_gold['Name'] / wint_combo_gold['Total Gold'] *100
wint_combo_gold
```

```
In [ ]:
wint_country_group_df = wint_combo_gold.groupby(['NOC'])

#Create DataFrame
wint_country_df = pd.DataFrame()
wint_country_df['Gold'] = wint_country_group_df['Name'].sum()
wint_country_df['Total Gold'] = wint_country_group_df['Total Gold'].sum()
wint_country_df['Gold%'] = wint_country_group_df['Gold %'].mean()
wint_country_df['latitude'] = wint_country_group_df['latitude'].mean()
wint_country_df.sort_values('Gold%', ascending = False)
```

Out[5]:

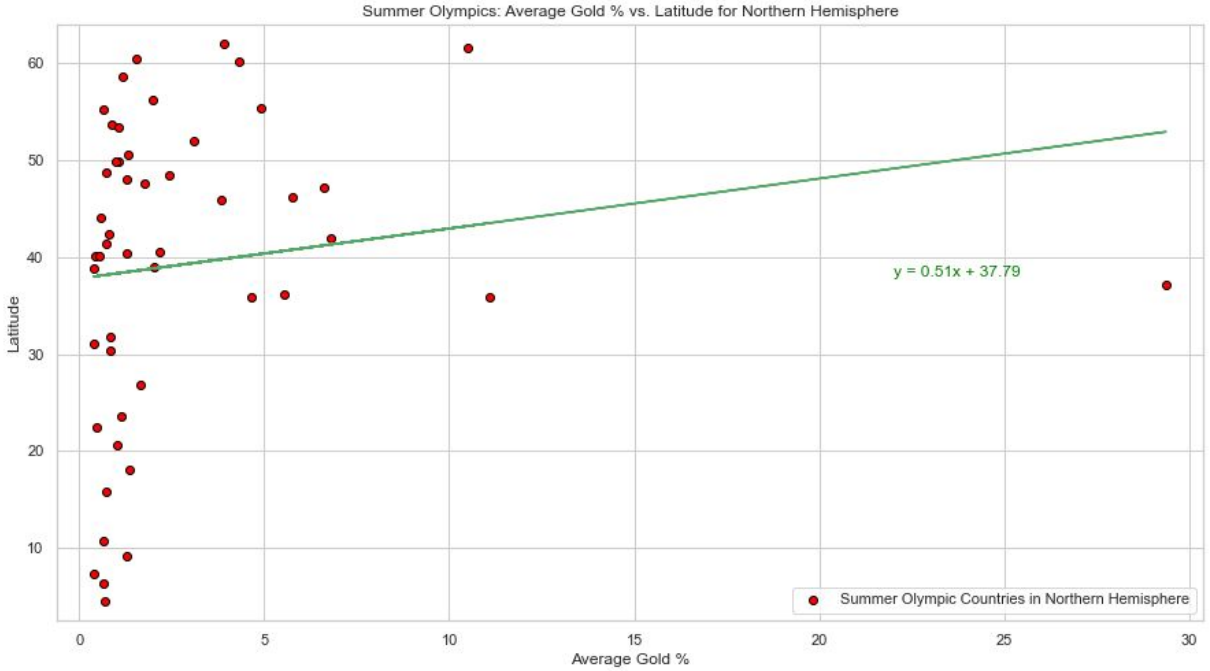
	Year	NOC_x	Season	Event	latitude	Name	gold	silver	bronze
70	1924	AUS	Winter	Alpinism Mixed Alpinism	-25.274398	1	1	0	0
99	1924	AUT	Winter	Figure Skating Mixed Pairs	47.516231	2	1	0	0
100	1924	AUT	Winter	Figure Skating Women's Singles	47.516231	1	1	0	0
248	1924	CAN	Winter	Ice Hockey Men's Ice Hockey	56.130366	9	1	0	0
410	1924	FIN	Winter	Speed Skating Men's 1,500 metres	61.924110	3	1	0	0
...	...	...	...	...	...	...	...	...	...
68933	2014	USA	Winter	Freestyle Skiing Men's Slopestyle	37.090240	4	1	1	1
68935	2014	USA	Winter	Freestyle Skiing Women's Halfpipe	37.090240	4	1	0	0
68964	2014	USA	Winter	Snowboarding Men's Slopestyle	37.090240	3	1	0	0
68966	2014	USA	Winter	Snowboarding Women's Halfpipe	37.090240	3	1	0	1
68967	2014	USA	Winter	Snowboarding Women's Slopestyle	37.090240	4	1	0	0

Out[10]:

NOC	Gold	Total Gold	Gold%	lati
NOR	111	620	21.099647	60.47
USA	96	636	16.568427	37.09
RUS	49	357	14.258778	61.52
SWE	50	499	12.567981	60.12
FIN	42	447	12.009879	61.92
LIE	2	17	11.764706	47.16
AUT	59	602	11.537638	47.51
CAN	62	564	9.363375	56.13
ITA	37	452	9.218378	41.87
ESP	1	12	8.333333	40.46
FRA	31	464	7.995078	46.22

Question: *Does the country's latitude impact their all-time average gold medal win percentage?*

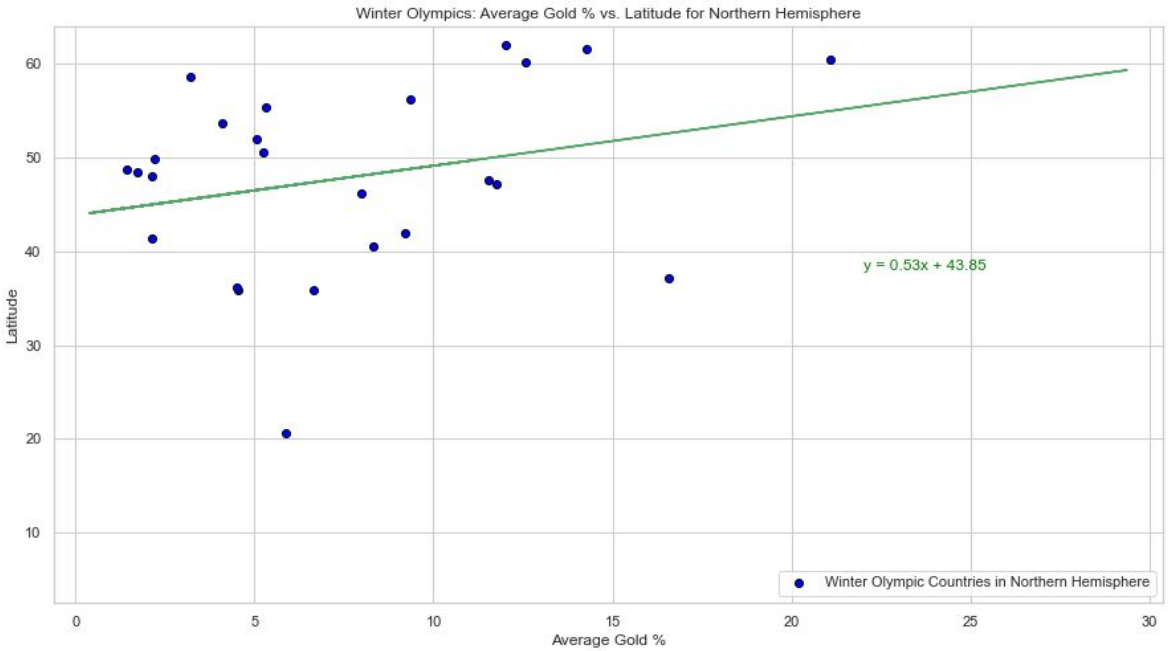
Summer Olympics: Average Gold % vs. Latitude for Northern Hemisphere



Observations

1. Majority of countries fall within the 1-2% regardless of latitude.
2. US has a much larger average Gold % than the rest of the World, with no other country breaking 12%.
3. The r value is 0.151 showing a weak positive correlation when we were really expecting a negative correlation.

Winter Olympics: Average Gold % vs. Latitude for Northern Hemisphere



Observations

1. Relative to Summer, countries are on much higher latitudes, with all but one above 30 degrees.
2. Relative to Summer, a much more balanced spread, with a majority of countries falling between 5 and 15%
3. With an r value of 0.267, the correlation of Average Gold % to Latitude is stronger than summer, but still relatively weak

Olympics Summary

Findings

- Countries near the equator with small populations are less likely to participate in the winter olympics
- Percentage of olympic athletes competing in summer is close to 100% when you drop below the latitude of 30 degrees
- Northern countries tend to participate more in winter olympics but that isn't always correlated with success
- Low correlation between performance and latitude for both total medals and gold medals
- For Summer, actually found the opposite correlation than we were expecting for both total medals and average percentage of gold medals

Conclusion

- With there being low correlation between performance and latitude, we were unable to reject our null hypothesis that changing latitude would not impact performance of countries at the Summer and Winter Olympics.

Next Steps

Difficulties:

1. There were too many variables that we were trying to capture which made it difficult to plot. When taking into accounts the various variables including season, country, location, year and count of participation, it became data overload.
2. With abundance of data, wanted to ask a multitude of questions. Challenging to focused on scope.
3. Weakness of summary variables of sums and averages across time. Don't tell the full picture
4. Unable to control for larger countries that span many latitudes. Limited by using only national capital latitude.

Follow-up questions:

1. Evolution of Sports- How has the participation in specific sport categories changed over time? Is it the same for both male and female athletes? Has the median age changed?
2. How will performance results change with the addition of 2018 and 2020 Olympic results?
3. Did COVID-19 impact the participation of countries in the 2020 Olympic Games?
4. How does does population size and economic status impact gold medal winnings?