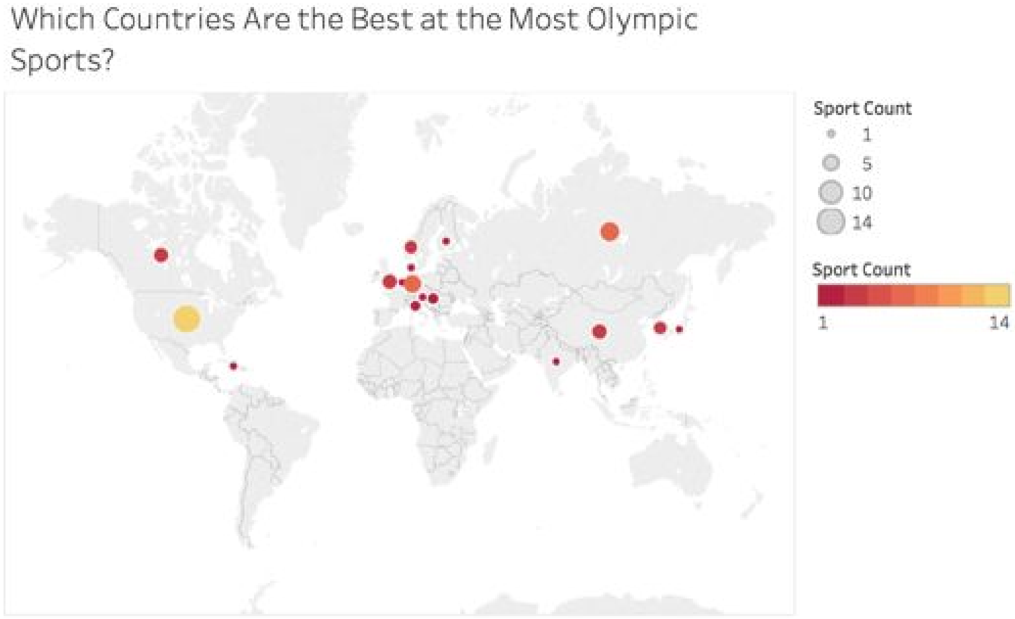
Data Visualization
Final Project
Edward Owens
Due 5/16/19

Introduction

The dataset visualized in this report, which came from Kaggle, contains information about every athlete who has ever competed at the Summer or Winter Olympics. For each athlete, the table has information about which Olympics they competed in, their home country, sport, event, and which medal they won if any. There is one row for each unique athlete-event combination, so some athletes have many rows. Michael Phelps, for example, has 30 rows. I also used population data from the United Nations. This dataset has population estimates for each country in every year from 1950 - 2015. I chose these dataset because I am a passionate fan of all sports, but the Olympics in particular, as I competed in Track and Field throughout High School and College. In this report, I explore which countries have achieved the most success overall and in certain sports. I also explore the performance of certain outstanding athletes. The population dataset was included in order to investigate which countries outperform on a per-capita basis.

Data Summary

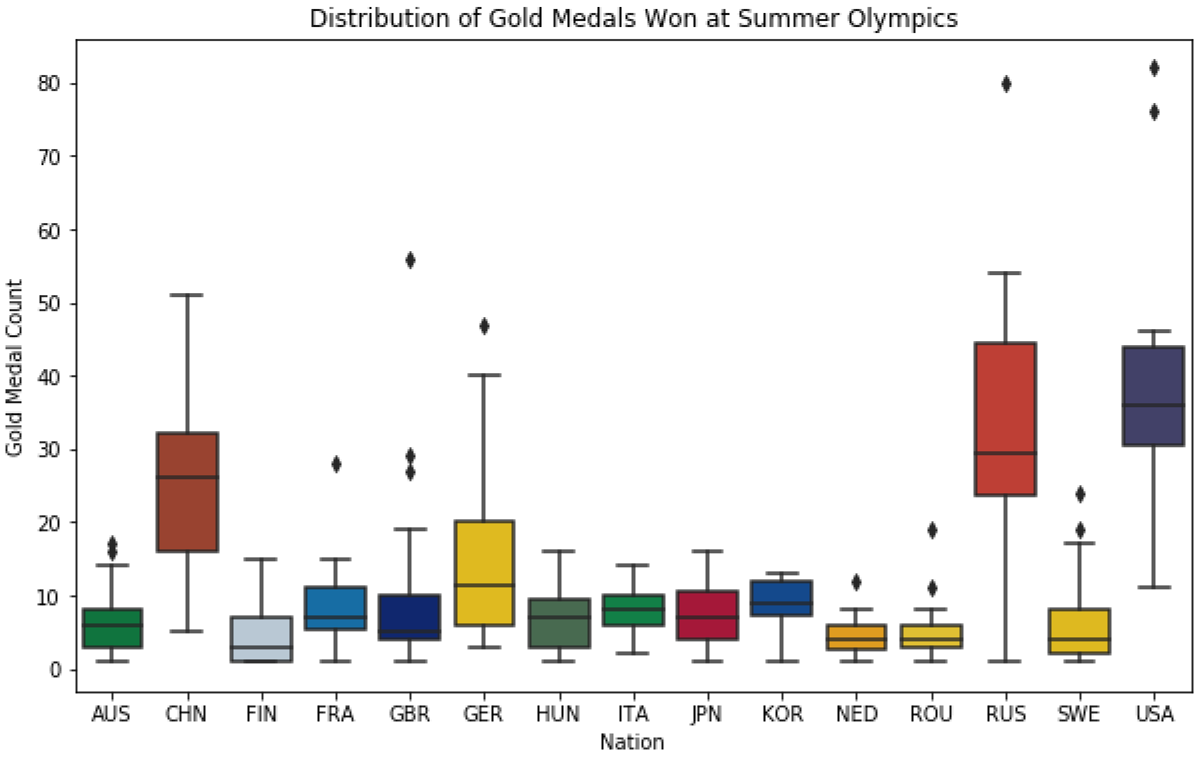
Let's start by exploring a very basic question: Which countries are the best at the most sports? To answer this, I've computed the number of gold medals won by each country in each sport that has appeared in at least 2 Olympic Games. I then determined which country had the most gold medals in each sport and tallied up the number of sports for each country. This bubble map displays the results:



Which Countries Are the Best at the Most Olympic Sports?

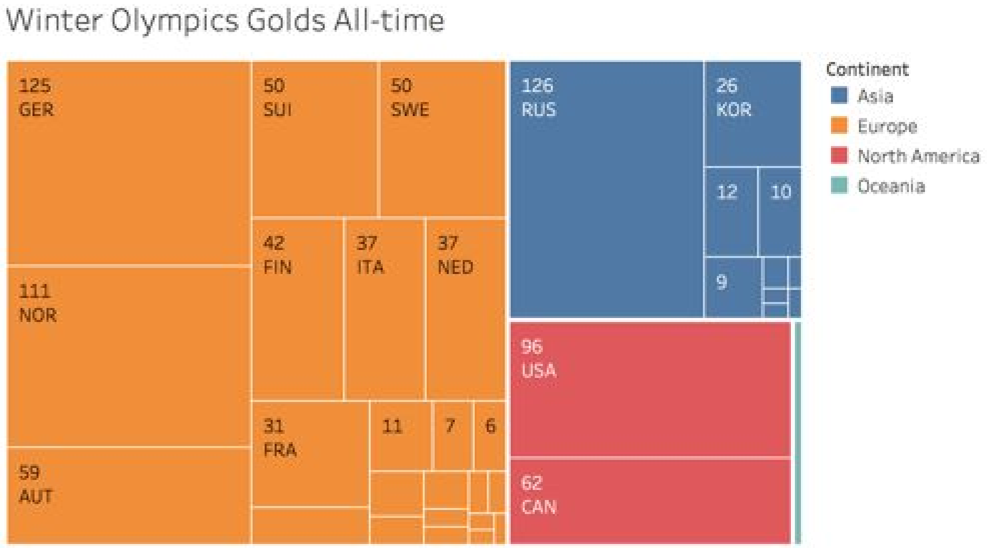
The US leads the way with 14 sports, Russia/Soviet Union is second with 7 sports, and Germany is third with 6 sports. Relative to its population, Norway is particularly impressive with 3 sports. Europe has the largest number of countries on this plot, several of which are best in just 1 sport. With 21 sports total, it also edges out North America’s 18. Only 3 continents are represented, each of which have a minimum of 16 sports.

Focusing on just the Summer Olympics now, let’s see what the distribution of Gold Medals won at a single Olympics looks like for some of the countries from the last plot. A boxplot is the perfect way to visualize this:



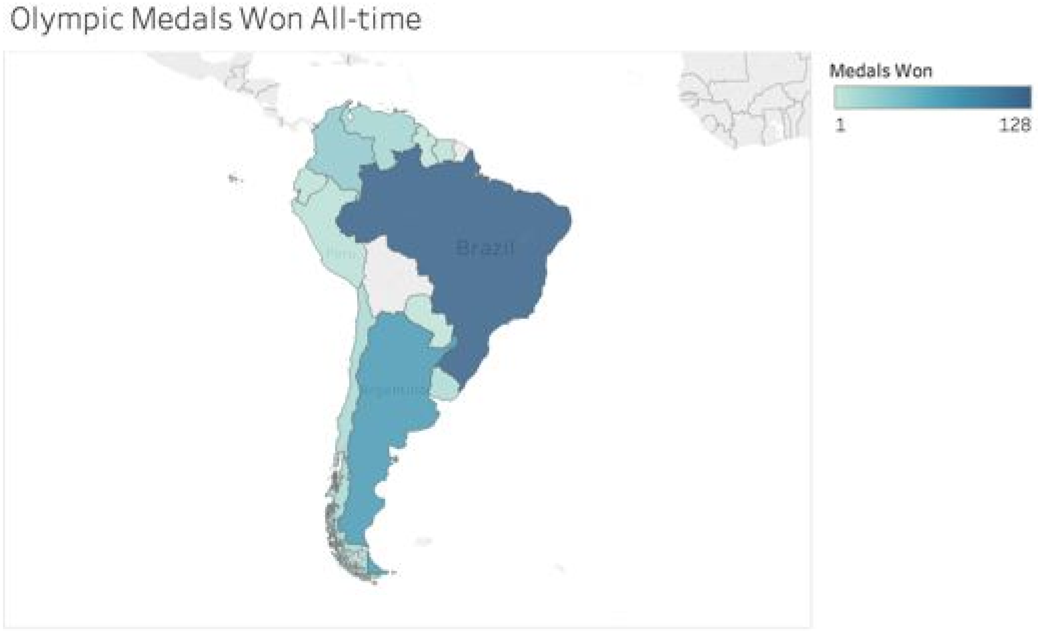
The same countries stand out again: The US, Russia, China, and to some extent Germany. Parts of Russia's distribution are actually higher than the USA's, although the US has the highest median. Most countries have a somewhat right-skewed distribution, although there is some diversity. Italy, for, example, has an almost perfectly symmetric distribution, while Korea's is somewhat left skewed.

Let's see if different countries stand out in the Winter Olympics. Here is a treemap of All-time Winter Olympic Gold Medals won by country grouped and color-coded by continent:



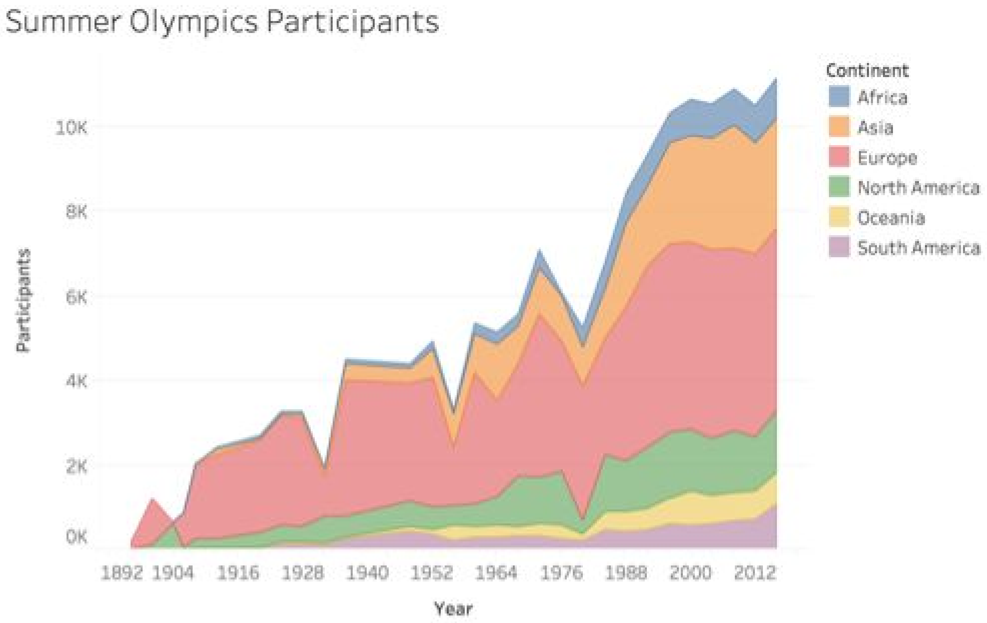
Europe clearly leads the way here, which makes sense given its climate and cultural ties to winter sports. However, it is Russia, which I've grouped with Asia, that narrowly edges out Germany for most Golds. Beyond those two, though, we start to see some different countries than the ones that dominate the Summer Olympics. The Scandinavian trio of Norway, Sweden, and Finland all appear, as does Canada. Apparently having cold weather helps your country succeed in the Winter Olympics!

Let's now focus on a part of the world that has not experienced tremendous Olympic success: South America. I've calculated the total number of Olympic Medals won all-time for all countries on the continent and visualized it with this choropleth:



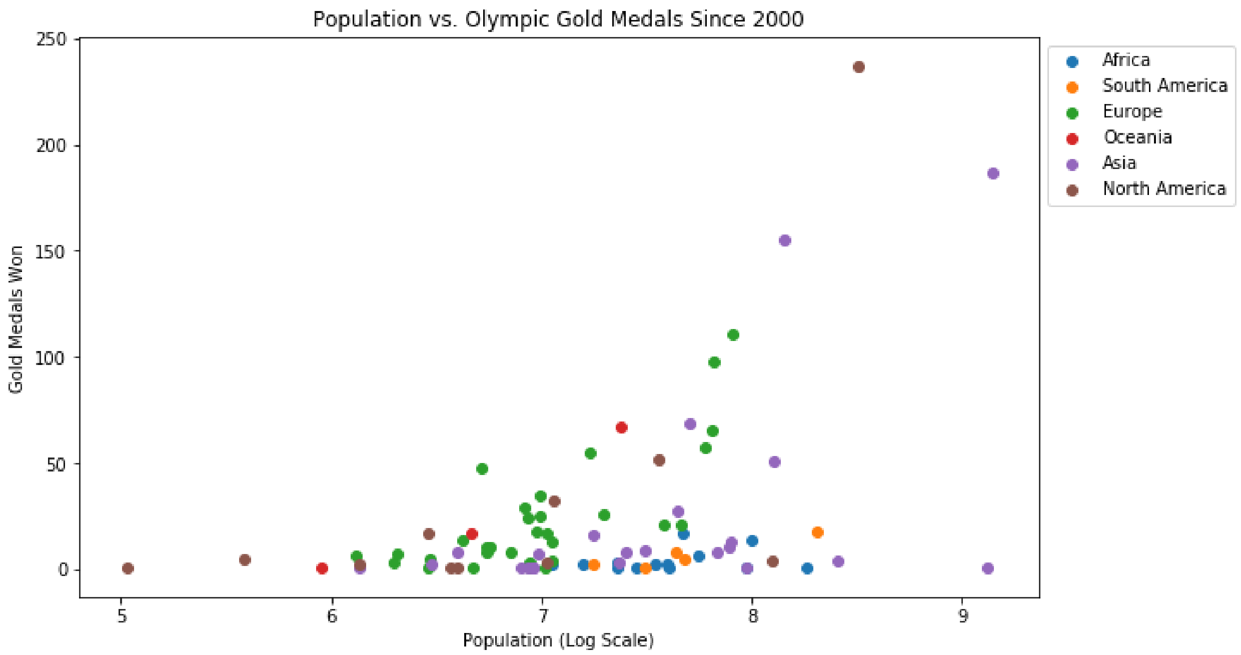
Unsurprisingly, Brazil, the most populous country on the continent, has won the most Olympic medals. Argentina and Colombia also stand out among the rest.

Let’s continue the regional trend and look at which continents the Summer Olympic athletes have come from over the history of the games. This streamplot displays the breakdown:



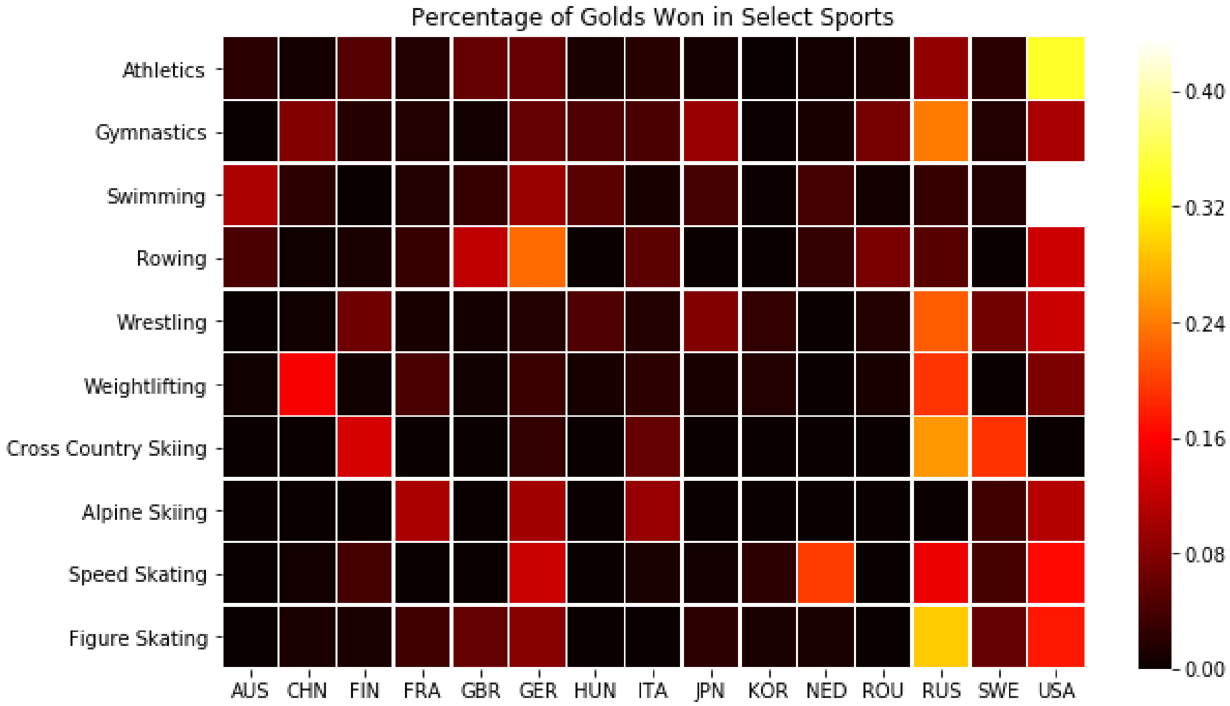
In the early years of the Olympics, nearly all the athletes hailed from Europe with some from North America. Almost no Asians or Africans competed until the 1940s. The number of Asian athletes has grown enormously since then, and it is now the second highest among the continents. However, it’s South America that appears to be the fastest growing region, at least since 2012. An interesting aside: it’s fascinating to see the effect that the US-led boycott of the 1980 Olympics had on the number of participants: a decrease of nearly 2000 athletes from a variety of countries.

So far we’ve looked at which countries have performed well, but let’s examine one factor that contributes to that success: population. Below, I’ve created a scatterplot of country population as of 2015 (on a log scale to reduce distortion) vs. Olympic Gold medals won since 2000:



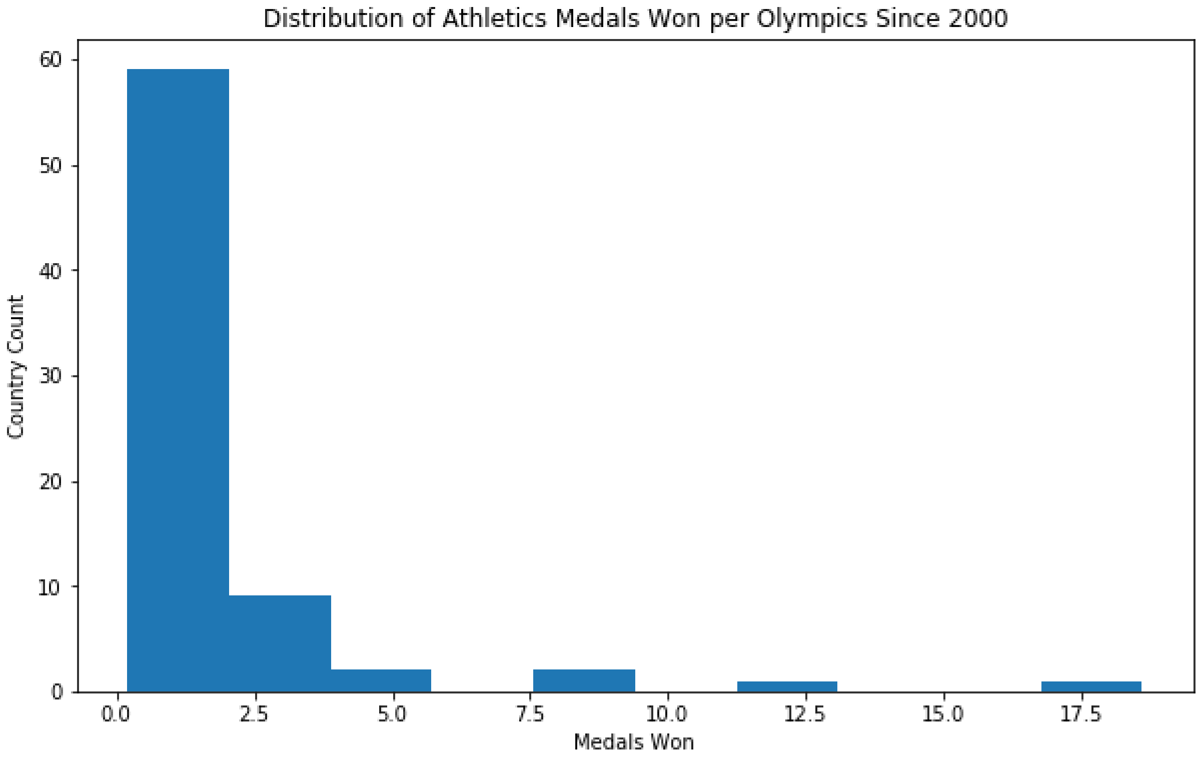
There is a clear exponential trend to this scatterplot: each additional unit of (log) population contributes more gold medals than the last, on average. Countries on certain continents appear to be better at translating their population into gold medals than countries from other continents. European countries, for example, appear higher up on the plot than, say, African countries. There are likely a number of factors at play here: European countries tend to have stronger cultural ties to Olympic competition than African countries and also are wealthier and stabler, on average, which means their athletes can make a living training and competing without needing other sources of income.

Let’s now turn our attention to some sport-specific breakdowns. I’ve taken the 15 countries from the boxplot earlier and calculated the percentage of all Gold medals that these countries won for some of the traditional Olympic sports. I’ve then visualized the results with this heatmap:



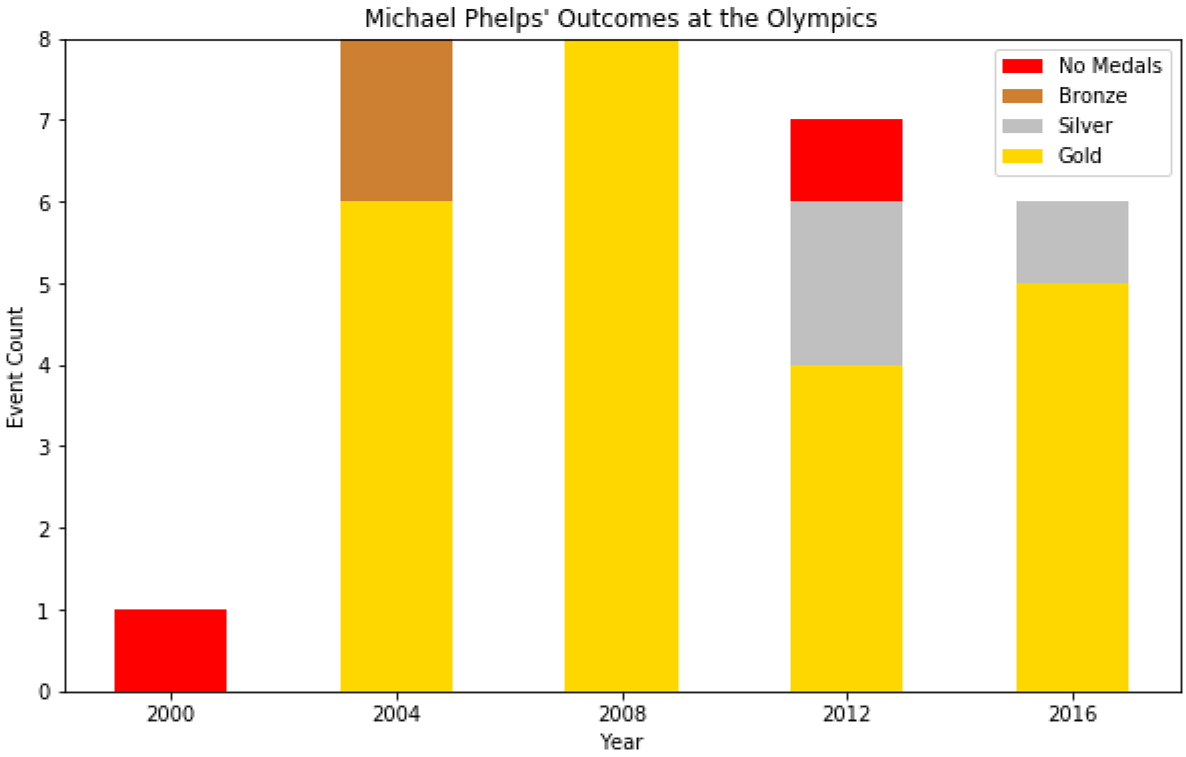
Let’s use one square as an example: The USA square for swimming is basically white, meaning the USA has won around 45% of all Gold medals ever awarded in swimming. This is the highest country-sport combination on this heatmap. It is interesting to see where certain countries’ strengths lie. The Netherlands, for example, are extremely good at Speed Skating, but barely register in the other sports. Germany is strongest in rowing, China is good at Weightlifting. Russia is good at many sports, but Figure Skating is its strongest.

I mentioned earlier that I competed in Track and Field (or Athletics in Olympics parlance), so I would be remiss if I didn’t include at least one chart about that sport. Athletics awards the most medals of any sport and has the most competitors, so I figured the distribution of Golds across countries might be interesting. I calculated the average number of Athletics medals won per Olympics for any country with at least one medal since 2000 and created a histogram to visualize the results:



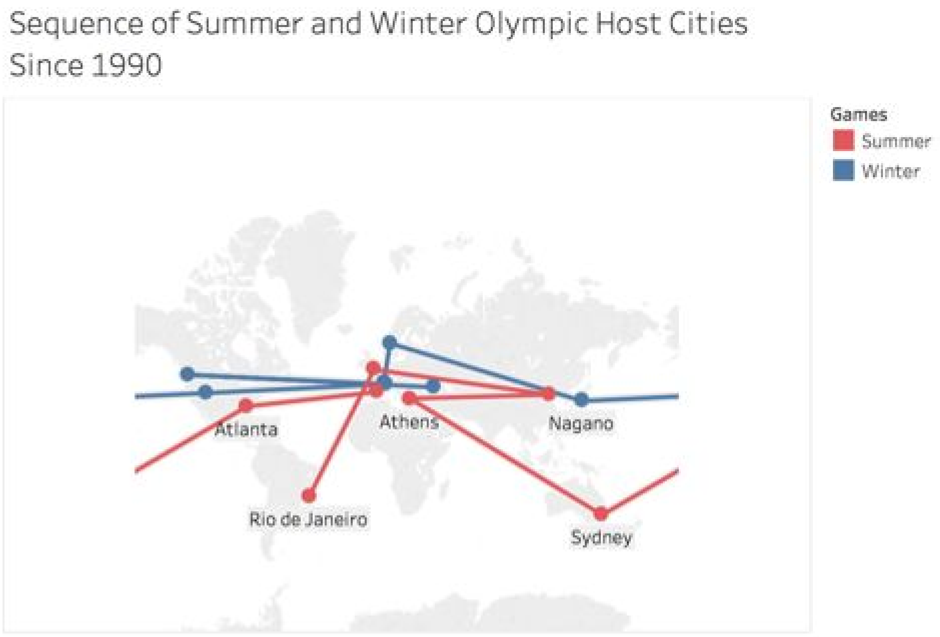
The result is a highly right-skewed distribution. Nearly 60 countries average 0-2 Athletics medals per Olympics! On the high end is the US, which averages around 18 medals per Games. Russia, Kenya, and Jamaica also all average more than 7 medals per Games. Kenya is a great example of an African country that has managed to create the infrastructure to enable its population to achieve Olympic success.

Let's now drill down to the finest level of detail that we can: an individual athlete. Perhaps no athlete in the history of the Olympics has been as closely associated with the Games than Michael Phelps. He made his first Olympics in 2000 at age 15 and over the next 16 years, won an astonishing 23 Gold medals, the most ever for a single athlete. I've created a bar chart showing his results for each Olympics:



The year that obviously stands out is 2008, when he spectacularly won Gold in all 8 events he entered. What people might forget is that he attempted the same feat in Athens 4 years earlier and came fairly close, with 2 bronzes his only blemishes. He battled personal struggles between 2008 and 2012, and was not quite at his peak in London. He failed to medal in an event for only the second time in his career, and won “only” 4 Golds. The Rio games 4 years later were his redemption: 5 Golds in 6 events was his second highest conversion rate.

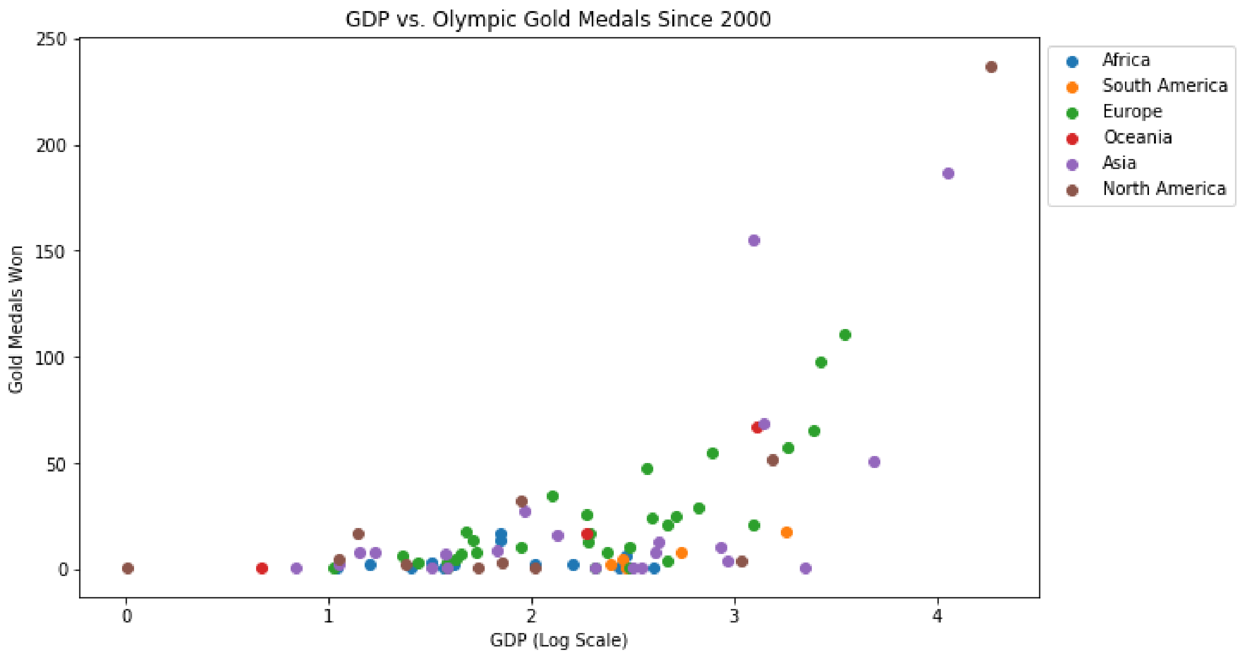
I’ll conclude this summary with a completely different plot that has nothing to do with winning medals. I’ve plotted the sequence of Summer and Winter Olympics host cities since 1990 using a connection map:



Unsurprisingly, the Winter Olympics have stayed in the Northern hemisphere, although Nagano is more southern than several Summer Olympic hosts. The Summer games have hit 5 of the 6 inhabited continents since 1990, with South America and Australia hosting for the first time during that span.

Storyline

Perhaps even more revealing than looking at how a country's population is tied to its Olympic success is considering how a country's GDP impacts it. The following scatterplot shows country GDP as of 2017 (on a log scale to reduce distortion) vs. Olympic Gold medals won since 2000:



At first glance this plot looks very similar to the population vs. gold medals plot as there is the same exponential trend. However, there are some differences in how the regions compare to one another. Unlike in the population plot, here the African countries don't show up as low outliers. For their GDP, they generally fall in line with the trend. In fact, only South America really stands out as being considerably below the trend. In this plot, Russia stands out as the largest positive outlier. They have won roughly twice as many gold medals as the next closest country among countries with equal or lesser GDPs. My hypothesis is that residuals from the trend reflect a country's systemic commitment to Olympics success (beyond just financial commitment). Russia has always placed a very high value on Olympic achievement, often going too far. For example, during the 2014 Winter Olympics, which they hosted in Sochi, Russia set up an elaborate system to evade anti-doping authorities. The result was a spectacularly successful Games, at least until the charade came falling down a few years later. On the opposite side, the South American countries simply don't seem to care as much about Olympic success. There are other sports, namely soccer, that those countries have poured resources into and excelled at.

Conclusion

Many of the plots presented here tell the same story: Both the Summer and Winter Olympic games have historically been dominated by countries from Europe, North America, and Asia. No countries outside these continents can claim to be the best at any single Olympic sport, as we saw in the bubble map. The Winter Olympics treemap tells an even more concentrated story: Australia is the only country to have ever won an Olympic gold medal outside those three continents.

Certain countries have been highlighted on multiple times as being among the most dominant. The US in particular, has led the way in this respect. With 14 sports claimed, it has double that of the next closest country. It also has the most Summer Olympics golds all-time and is 4th in Winter Olympics golds. It has been the dominant country in two of the most marquee Olympic sports: Track and Field and Swimming. Russia has also been impressively dominant: second in Summer Olympics golds, first in Winter golds, and the dominant country in one of the marquee Winter Olympic sports, figure skating. Other countries with high levels of achievement include China, Germany, and Norway, specifically in the Winter Olympics.

Among the countries on continents without a history Olympics success, Brazil, Kenya, and Australia stand out. There are many factors that contribute to a country's level of Olympic success. A few that I have posited are the country's traditional ties to Olympic competition, the country's population, and the country's wealth.

Code Appendix (Github link: https://github.com/erowens/dataviz_project)

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

olympics = pd.read_csv('olympics/athlete_events.csv', index_col=0)


# Bubble Map

sport_games_cts = olympics.groupby('Sport').nunique().Games
sport_games_cts = sport_games_cts[sport_games_cts>=2]
sport_index = olympics.set_index('Sport')
olympics_refined = sport_index.loc[sport_games_cts.index]
golds = olympics_refined[olympics_refined.Medal=='Gold']
sport_golds = golds.groupby(['Sport', 'NOC']).count().Medal

sports = []
top_countries = []
for sport in sport_golds.index.levels[0]:
    sport_gold_count = sport_golds.loc[sport]
    sports.append(sport)
    top_countries.append(sport_gold_count.idxmax())

top_countries = ['RUS' if country == 'URS' else country for country in top_countries]
country_sports = pd.Series(sports, index=top_countries)
countries = []
```

```
sport_strings = []
sport_cts = []
for country in country_sports.index.unique():
    country_sport = country_sports.loc[country]
    if type(country_sport) == str:
        sport_strings.append(country_sport)
        sport_cts.append(1)
    else:
        all_sports = ', '.join(country_sport.values)
        sport_strings.append(all_sports)
        sport_cts.append(country_sport.shape[0])
    countries.append(country)

final_info = pd.DataFrame([sport_cts, sport_strings], columns=countries, index=['count',
'string']).transpose()
final_info.to_csv('country_cts.csv')


# Boxplot

only_summer_golds = olympics[(olympics.Medal == 'Gold') & (olympics.Season == 'Summer')]

class smart_dict(dict):
    def __missing__(self, key):
        return key

only_summer_golds['NOC'] = only_summer_golds.NOC.map(smart_dict({'URS':'RUS',
'GDR':'GER'})).copy()
country_gold_cts = only_summer_golds.groupby(['NOC', 'Games']).nunique().Event
top15 = country_gold_cts.groupby(['NOC']).sum().sort_values(ascending=False).iloc[:15]
top15_golds = country_gold_cts.loc[top15.index].reset_index()

nation_colors = ['#00843D',  # AUS
                 '#aa381e',  # CHN
                 '#b5c9d8',  # FIN
                 '#0072bb',  # FRA
                 '#001F7E',  # GBR
                 '#FFCE00',  # GER
                 '#436F4D',  # HUN
                 '#009246',  # ITA
                 '#bc002d',  # JPN
                 '#0047A0',  # KOR
                 '#FFA500',  # NED
```

```
            '#FCD116',  # ROU
            '#D52B1E',  # RUS
            '#fecc00',  # SWE
            '#3C3B6E',  # USA
           ]

plt.figure(figsize=(10,6))
sns.boxplot(data=top15_golds, x='NOC', y='Event', palette=nation_colors)
plt.title('Distribution of Gold Medals Won at Summer Olympics')
plt.xlabel('Nation')
plt.ylabel('Gold Medal Count')
plt.show()


# Histogram

athletics = olympics[(olympics.Sport == 'Athletics') & (olympics.Year >= 2000)]
athletics_medals = athletics[1-athletics.Medal.isna() == 1]
country_athletics_medals = athletics_medals.groupby(['NOC', 'Year']).Event.nunique()

plt.figure(figsize=(10,6))
plt.hist(country_athletics_medals.groupby(['NOC']).sum()/5)
plt.title('Distribution of Athletics Medals Won per Olympics Since 2000')
plt.xlabel('Medals Won')
plt.ylabel('Country Count')
plt.show()


# Barplot

phelps = olympics[olympics.Name == 'Michael Fred Phelps, II']
phelps.fillna('None', inplace=True)
phelps_byyear = phelps.groupby(['Year', 'Medal']).Event.nunique()

medal_ct = []
years = []
medals = []
for year in phelps_byyear.index.levels[0]:
    for medal in phelps_byyear.index.levels[1]:
        if medal in phelps_byyear.loc[year]:
            medal_ct.append(phelps_byyear.loc[year].loc[medal])
        else:
            medal_ct.append(0)
        years.append(year)
        medals.append(medal)

full_phelps = pd.DataFrame({'Year':years, 'Medal_ct':medal_ct}, index=medals)

plt.figure(figsize=(10,6))
bar_Gold = plt.bar(full_phelps.loc['Gold'].Year, full_phelps.loc['Gold'].Medal_ct, color='gold',
width=2)
bar_Silver = plt.bar(full_phelps.loc['Silver'].Year, full_phelps.loc['Silver'].Medal_ct,
                   bottom=full_phelps.loc['Gold'].Medal_ct, color='Silver', width=2)
bar_Bronze = plt.bar(full_phelps.loc['Bronze'].Year, full_phelps.loc['Bronze'].Medal_ct,

bottom=full_phelps.loc['Gold'].Medal_ct.values+full_phelps.loc['Silver'].Medal_ct.values,
                  color='#cd7f32', width=2)
bar_None = plt.bar(full_phelps.loc['None'].Year, full_phelps.loc['None'].Medal_ct,

bottom=full_phelps.loc['Gold'].Medal_ct.values+full_phelps.loc['Silver'].Medal_ct.values+
                full_phelps.loc['Bronze'].Medal_ct.values, color='Red', width=2)

plt.xticks([2000, 2004, 2008, 2012, 2016])
plt.legend((bar_None[0], bar_Bronze[0], bar_Silver[0], bar_Gold[0]), ('No Medals', 'Bronze',
'Silver', 'Gold'),
          bbox_to_anchor=(1,1))
plt.title("Michael Phelps' Outcomes at the Olympics")
plt.xlabel('Year')
plt.ylabel('Event Count')
plt.show()


# Heat Map

sport_sample = ['Athletics', 'Gymnastics', 'Swimming', 'Rowing', 'Wrestling', 'Weightlifting',
               'Cross Country Skiing', 'Alpine Skiing', 'Speed Skating', 'Figure Skating']
sport_binary = [True if sport in sport_sample else False for sport in olympics.Sport]
select_sports = olympics[sport_binary & (olympics.Medal=='Gold')]
select_sports['NOC'] = select_sports.NOC.map(smart_dict({'URS':'RUS', 'GDR':'GER'}))
country_full_cts = select_sports.groupby(['NOC', 'Games',
'Sport']).Event.nunique().reset_index()
top_countries = top15_golds.NOC.unique()
country_sport_cts = country_full_cts.groupby(['Sport','NOC']).Event.sum()

full_data = []
for sport in sport_sample:
    sport_data = country_sport_cts.loc[sport]
    tot = sum(sport_data)
    sport_list = []
    for country in top_countries:
        if country in sport_data.index:
            country_sport = sport_data.loc[country]
            sport_list.append(country_sport/tot)
        else:
            sport_list.append(0)
    full_data.append(sport_list)

matrix = pd.DataFrame(full_data, index=sport_sample, columns=top_countries)

plt.figure(figsize=(10,6))
ax = sns.heatmap(matrix, linewidth=0.5, cmap='hot')
plt.title('Percentage of Golds Won in Select Sports')
# plt.xlabel('Country')
# plt.ylabel('Sport')
plt.show()


# Choropleth

south_american = ['ARG', 'BOL', 'BRA', 'CHI', 'COL', 'ECU', 'GUY', 'PAR',
                 'PER', 'SUR', 'URU', 'VEN']
south_american_data = olympics[[True if country in south_american else False for country in
olympics.NOC]]
south_american_data.fillna('None', inplace=True)
south_american_data = south_american_data[south_american_data.Medal != 'None']
sa_medal_cts = south_american_data.groupby(['NOC', 'Games',
'Medal']).Event.nunique().reset_index()
sa_medal_cts.groupby(['NOC']).Event.sum().to_csv('south_america.csv')


# Connection Map

host_cities = olympics.groupby('Games').City.unique()
host_cities_full = pd.DataFrame([value.split() for value in host_cities.index.values],
columns=['Year', 'Games'])
host_cities_full['City'] = [array[0] for array in host_cities]
host_cities_full.sort_values(['Games', 'Year']).to_csv('connection.csv')
```

```
# Stream graph

regions = pd.read_csv('olympics/noc_regions.csv', index_col=0)
region_dict = {noc:regions.continent.loc[noc] for noc in regions.index}
olympics['continent'] = olympics.NOC.map(region_dict)
olympics[olympics.Season=='Summer'].groupby(['Year',
'continent']).Name.nunique().to_csv('stream.csv')


# Treemap

all_years = olympics[(olympics.Season=='Winter') & (olympics.Medal=='Gold')].groupby(['Year',
'continent', 'NOC']).Event.nunique()
all_years = all_years.reset_index()
all_years['NOC'] = all_years.NOC.map(smart_dict({'GDR':'GER', 'URS':'RUS'}))
all_years.groupby(['continent', 'NOC']).Event.sum().to_csv('treemap.csv')


# Scatterplot

population = pd.read_excel('world_pop2.xlsx', index_col=0)
pop_2015 = population[2015]
country_golds = olympics[(olympics.Year>=2000) & (olympics.Medal=='Gold')].groupby(['NOC',
'Games']).Event.nunique()
country_golds_tot = country_golds.reset_index().groupby('NOC').sum()

country_codes = {noc:regions.region.loc[noc] for noc in regions.index}
region_dict = {noc:regions.continent.loc[noc] for noc in regions.index}
continent_map = {'africa':'Africa', 'asia':'Asia', 'northam':'North America',
                'europe':'Europe', 'southam':'South America', 'oceania':'Oceania'}

country_golds_tot['country'] = country_golds_tot.index.map(country_codes)
country_golds_tot['continent'] = country_golds_tot.index.map(region_dict)
country_golds_tot.set_index('country', inplace=True)

country_gold_popn = country_golds_tot.join(pop_2015, how='outer')
country_gold_popn = country_gold_popn[(country_gold_popn.Event>0)]

plt.figure(figsize=(10,6))
axes = []
continents = []
for continent in country_gold_popn.continent.unique():
    if continent not in ['none', np.nan]:
        continent_golds = country_gold_popn[country_gold_popn.continent==continent]
        axes.append(plt.scatter(np.log10(continent_golds[2015]*1000), continent_golds.Event))
        continents.append(continent_map[continent])
plt.title('Population vs. Olympic Gold Medals Since 2000')
plt.xlabel('Population (Log Scale)')
plt.ylabel('Gold Medals Won')
plt.legend(axes, continents, bbox_to_anchor=(1,1))
plt.show()


# Storyline Scatterplot

gdp = pd.read_excel('gdp.xlsx', index_col=4)
country_gold_gdp = country_golds_tot.join(gdp, how='outer')
country_gold_gdp = country_gold_gdp[(country_gold_gdp.Event>0)]

plt.figure(figsize=(10,6))
axes = []
continents = []
for continent in country_gold_gdp.continent.unique():
    if continent not in ['none', np.nan]:
        continent_golds = country_gold_gdp[country_gold_gdp.continent==continent]
        axes.append(plt.scatter(np.log10(continent_golds.unGDP), continent_golds.Event))
        continents.append(continent_map[continent])
plt.title('GDP vs. Olympic Gold Medals Since 2000')
plt.xlabel('GDP (Log Scale)')
plt.ylabel('Gold Medals Won')
plt.legend(axes, continents, bbox_to_anchor=(1,1))
plt.show()
```