Machine Learning with Olympic Data

Members: Kathleen Campbell, Kate Yip, John Flexner, Devin Luka, Jahmarli Cohen

Source: Google Images

The Olympics:

- International sports competition that happens every 4 years.
- Countries all over the world compete against each other in various sports events/disciplines, both team and individually.
- Originated in Ancient Greece
- 4 Different Olympics: Summer, Winter, Paralympics, & Youth
- Top 3 winners in each sports competition are rewarded with medals:
    - Gold- 1st Place
    - Silver- 2nd Place
    - Bronze- 3rd Place

Source: https://olympics.com/en/sports/

Overview of the Olympic Data Sets

- Excel files separated into 5 different Olympic categories:
  - Athletes
  - Gender
  - Discipline/Sport
  - Teams by Event
  - Medals
- Summer Olympics Data
- Over 100 countries participated
- Team USA won the most medals.
- Approximately 90 countries won at least 1 medal.

Evolution of Medal Design

Source: https://www.nbcnewyork.com/news/sports/beijing-winter-olympics/the-history-of-all-modern-olympic-medals/3154180/

Medals Per Country- Total, Gold, Silver, and Bronze. Includes Rank by Performance & Rank by Total Medals.

Rank	NOC	Gold	Silver	Bronze	Total	Rank by Total
1	People's Republic of China	32	22	16	70	2
2	United States of America	25	30	22	77	1
3	Japan	21	7	12	40	5
4	Great Britain	15	18	15	48	4
5	Australia	15	4	17	36	6
6	ROC	14	21	18	53	3
7	Germany	8	8	16	32	7
8	France	6	10	9	25	9
9	Italy	6	9	15	30	8
10	Netherlands	6	8	9	23	10
11	Republic of Korea	6	4	9	19	11
12	New Zealand	6	4	5	15	12
13	Cuba	5	3	4	12	15
14	Hungary	4	5	3	12	15
15	Brazil	4	3	8	15	12

GDP (Billions of USD)

Rank (GDP)	Name of Country	GDP (Billions of USD)	Total Medals
1	United States of America	20940	77
2	People's Republic of China	14720	70
3	Japan	5065	40
4	Germany	3806	32
5	Great Britain	2708	48
6	India	2623	3
7	France	2603	25
8	Italy	1886	30
9	Canada	1643	14
10	Republic of Korea	1631	19

Pivot Table of Team Events by Country & their Total Medals.

	A	B	C
1	Countries	# of Team Events	Medals
2	United States of America	47	77
3	People's Republic of China	33	70
4	ROC	34	53
5	Great Britain	28	48
6	Japan	48	40
7	Australia	35	36
8	Germany	36	32
9	Italy	37	30
10	France	33	25
11	Netherlands	27	23
12	Republic of Korea	19	19
13	Brazil	25	15
14	New Zealand	13	15
15	Canada	30	14
16	Cuba	3	12
17	Hungary	14	12
18	Switzerland	12	12
19	Chinese Taipei	7	11
20	Ukraine	10	11

Purpose of Our Project:

- Question: Can we use the Olympics dataset to predict a country's medal count?
- Hypothesis: Medal count can be predicted with decent accuracy using the Olympics dataset
- We worked as a team to gather and select variables from the data that could be used to predict a country's medal count
- We then used these variables to construct visualizations of the data and machine learning to make predictions
- Finally, we used Statistics to evaluate our variables and final model

What Methods/ Tools did We Use?

- Python 3
  - Numpy
  - Pandas
  - Matplotlib
  - Scikit Learn
  - Style (Seaborn)
  - Geopandas*
  - Plotly

Along with Python 3, we also used various Excel tools.

This includes:

- Pivot Tables
- Pivot Charts
- Sorting
- Value Filters
- Label Filters
- Non Pivot Tables & Charts
- Statistical Analysis
- Data Analysis

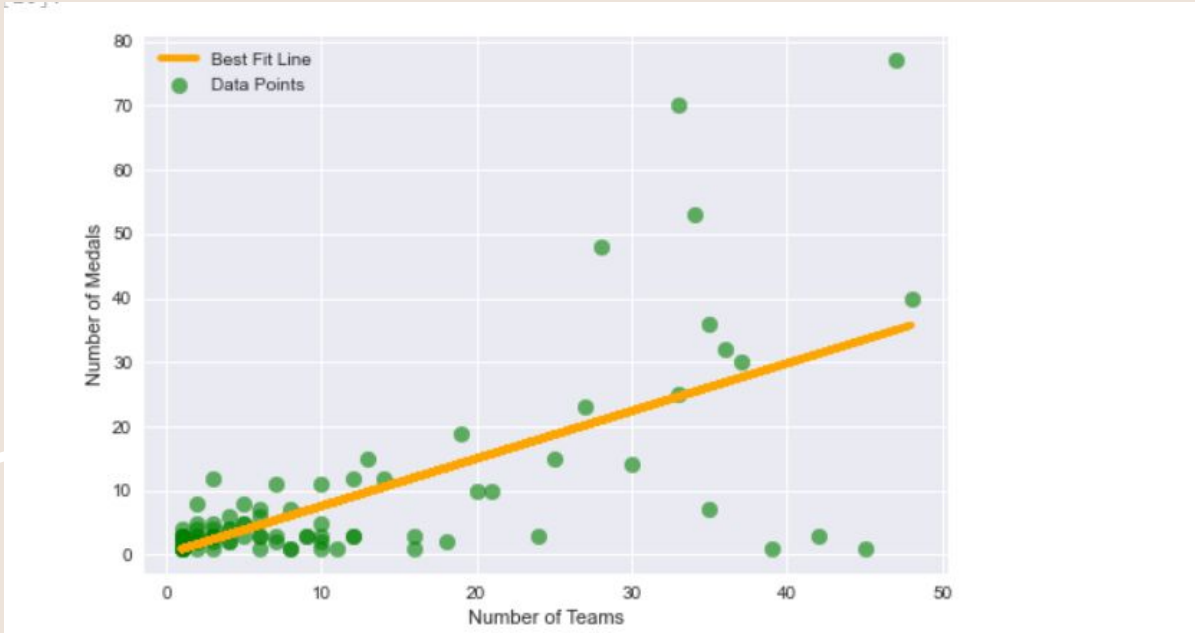
R Squared

- 73.777 % of the variation in number of medals a country wins can be explained by the number of athletes each country sends.
- 45.249 % of the variation in number of medals a country wins can be explained by the number of disciplines a country participates in.
- 67.532 % of the variation in number of medals a country wins can be explained by the GDP of a country
- 44.013 % of the variation in number of medals a country wins can be explained by the amount of team events a country participated in.
- 0.859343526903836, or about 86% of the variation in medal count can be explained by the model that includes all of the above variables

Machine Learning Model

Linear Regression

The predictive analysis used to give estimates of how to describe data and explain the relationship between two variables.



X & Y Variables

Independent Variables (x):

- Disciplines/Sport
- # of Teams (per country)
- GDP (Billions USD)
- Athletes (per country)

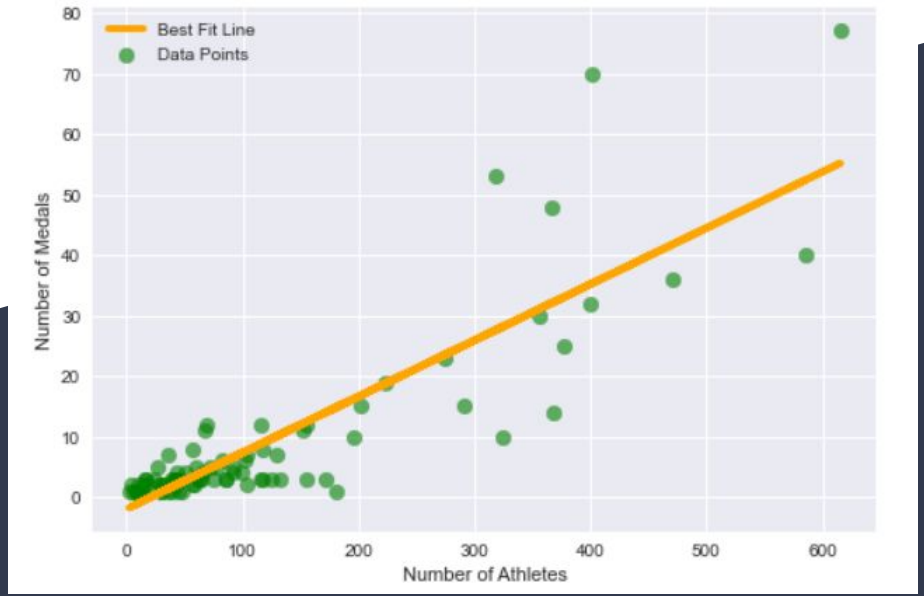
Dependent Variable (y):

- # of Medals (per country)

Athletes x Medal Count

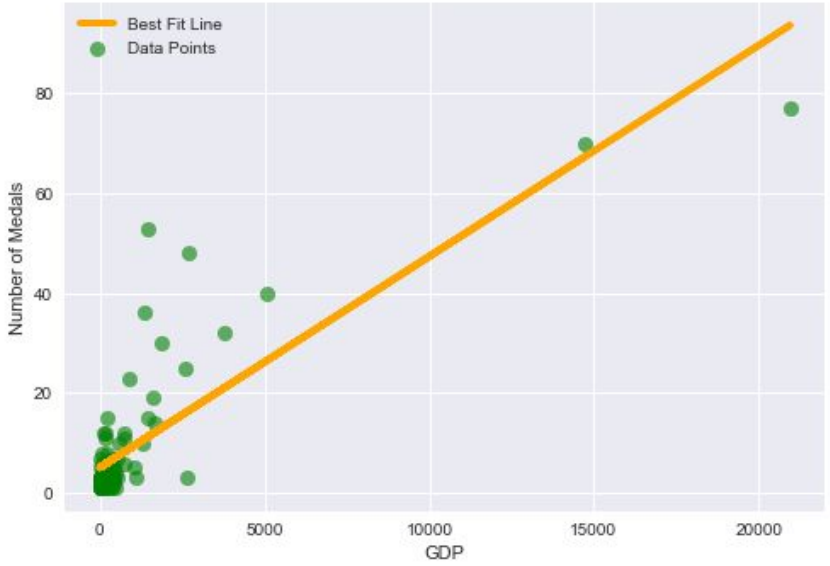
In [21]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from matplotlib.pyplot import style
from sklearn import linear_model

#Convert Medals and Athletes to columns
athletes_list = df['Number of Athletes'].tolist()
medals_list = df['Number of medals'].tolist()

#convert lists to numpy lists
x = np.array(athletes_list,dtype = np.float64)
y = np.array(medals_list,dtype = np.float64)

#Create linear regression object
medals_athletes = linear_model.LinearRegression()

#Train the model using the training sets
medals_athletes.fit(x.reshape(-1,1),y)

#get the regression line using the model
regression_line = medals_athletes.predict(x.reshape(-1,1))


#plot points
style.use('seaborn')
plt.scatter(x, y, label = 'Data Points', alpha = 0.6, color = 'g', s = 75 )
plt.plot(x, regression_line, label = 'Best Fit Line', color = 'orange', linewidth = 4)
plt.xlabel('Number of Athletes')
plt.ylabel('Number of Medals')
plt.legend()
```



GDP (Billions USD) – Per Country

```
#Convert Medals and Athletes to columns
Gdp_list = df['GDP'].tolist()
medals_list = df['Number of medals '].tolist()

#convert lists to numpy lists
x = np.array(Gdp_list,dtype = np.float64)
y = np.array(medals_list,dtype = np.float64)

#Create linear regression object
medals_Gdp = linear_model.LinearRegression()

#Train the model using the training sets
medals_Gdp.fit(x.reshape(-1,1),y)

#get the regression line using the model
regression_line = medals_Gdp.predict(x.reshape(-1,1))


#plot points
style.use('seaborn')
plt.scatter(x, y, label = 'Data Points', alpha = 0.6, color = 'g', s = 75 )
plt.plot(x, regression_line, label = 'Best Fit Line', color = 'orange', linewidth = 4)
plt.xlabel('GDP')
plt.ylabel('Number of Medals')
plt.legend()
```

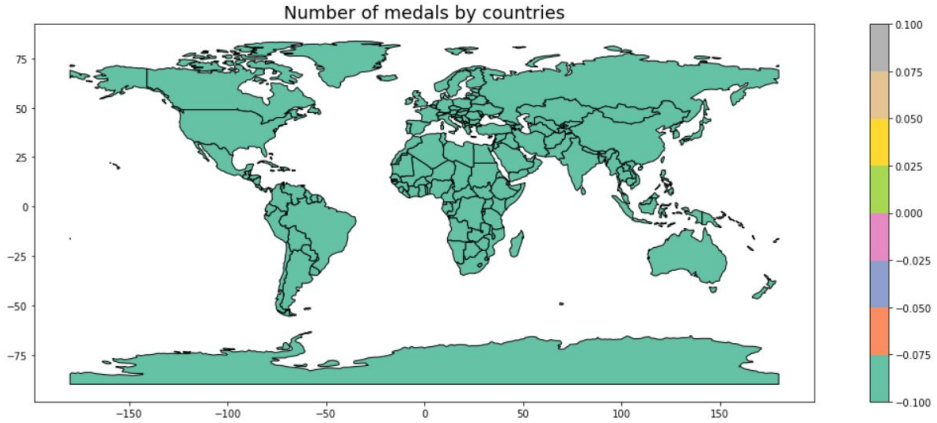


Geopandas

Geopandas is a Python module that works geospatial data and gives us a geographical visual of a world map based on longitude and latitude. Our team merged our collective data to create a geographical profile that displays the total number of medals (legend) by country.

```
import pandas as pd
import matplotlib.pyplot as plt
import geopandas as gpd
from geopandas import GeoDataFrame


gdf = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
gdf['medals']=0

fig, ax = plt.subplots(figsize=(20,7))

gdf.plot(column='medals',ax=ax,legend=True,cmap='Set2',edgecolor="black")
plt.title("Number of medals by countries", fontsize=18)
```

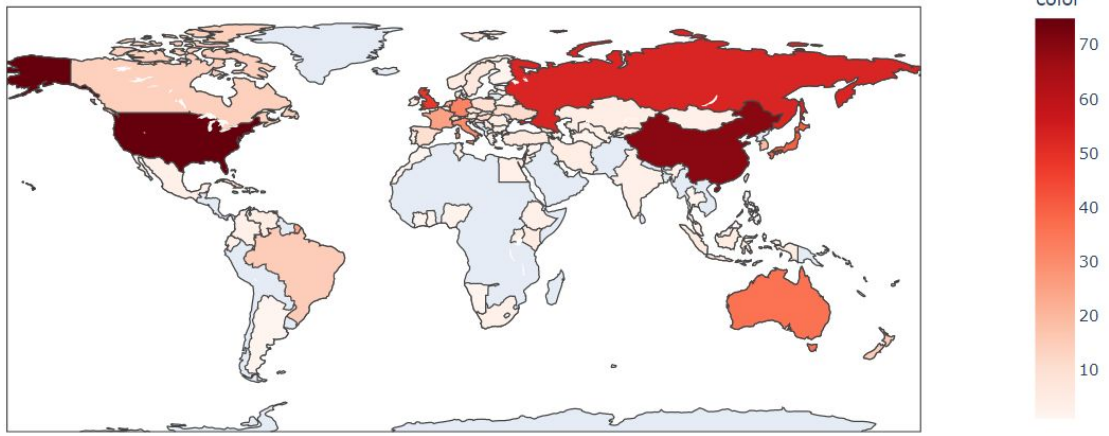
```
Text(0.5, 1.0, 'Number of medals by countries')
```



Plotly

```
import plotly.express as px
x_ath = df['Number of Athletes']
x_cod = df['Count of Discipline/Sport']
x_gdp = df['GDP']
x_team = df['Number of Teams']
y = df['Number of Medals']
plt.figure(figsize=(100,70))
px.choropleth(y, locations=df['Countries'],
                    locationmode='country names', color=y,
                    hover_data= ['Number of Medals'], range_color=[1,75],
                    color_continuous_scale='reds',
                    title='Number of Medals per Country')
```

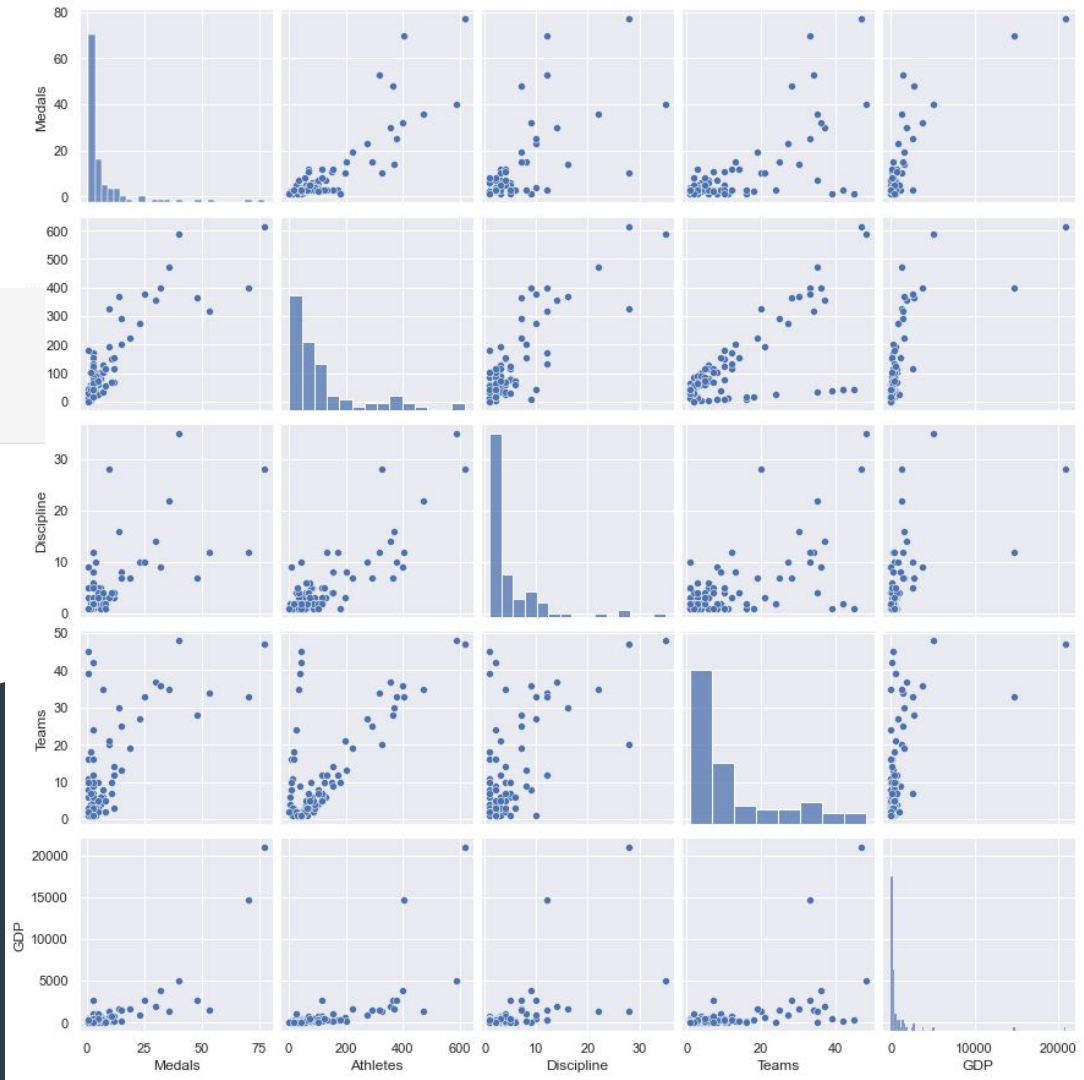
Number of Medals per Country



Pair Plot

```
# pairplot of our variables, each observation is a country
%matplotlib inline
import seaborn as sns; sns.set()
sns.pairplot(df_1);
```

A simple pair plot:
Diagonals show histograms for each variable while the outer plots are the scatterplots for each pair



Summary of Our Results

- From our analysis of the Linear Regression model, we were able to determine that:
  - Increasing the number of disciplines/sport would boost the medal count by 1.53558966 points.
  - For each additional athlete sent by a country the number of medals increases by 0.09280544 points.
  - For every added team event a country participated in the number of medals increased by 0.74130302
  - Increasing the GDP by one unit would would cause the medal count for a country to increase by 0.00422817
  - R Squared: 0.859343526903836, about 86% of the variation in medal count can be explained by the model

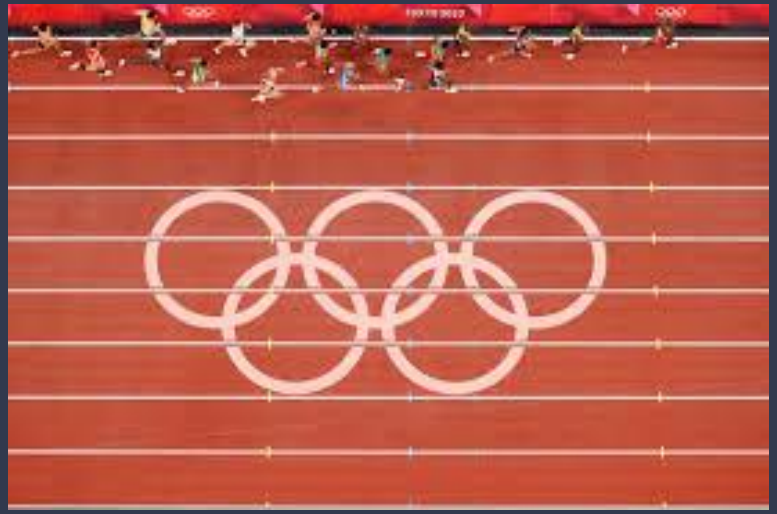
Interesting Notes/ Future Areas of Exploration

- **Interesting Notes:**
  - For simplicity, we used data from medal winning countries
  - Regression for Number of Athletes and GDP resulted in an R-squared of 0.84, almost as significant as the overall model
- **Future Areas of Exploration:**
  - Age - using age to determine how varying experience levels could affect medal count.
  - Gender - What role does gender play in medal count? Who brings home more medals and why?

Thank You!

Team Members: Kate Yip, Kathleen Campbell, John Flexner, Devin Luka, Jahmarli Cohen

Feb. 2022



Source: Google Images