#Machine learning model notes:

##Preprocessing:

Data was collected from multiple sources including webscraping from Olympedia.org, TradingEconomics, the OECD database, and the World Bank Database. The data was cleaned using excel and pandas.

The cleaned data was then merged into two tables. One table included information related to countries competing at the Olympics. The second table included the individual medal results for each Olympic event. The data was organized by year, discipline, gender, and results. Additionally, medal results were given a scoring range with gold equal to 3, silver equal to 2, and bronze equal to 1. The number were chosen to provide more weight to gold medals during the model training which focused on maximizing a higher medal “score”.

##Features:

###Features selected included:

* Medals won by each country from 1964-2016
* Country GDP per year from 1964-2016. GDP that was not able to be collected was filled using either the average GDP of the country in question given we had collected enough data to confidently apply the average to missing years, or the data was filled using the average of the entire dataset. The hypothesis is that richer countries are likely to outperform poorer countries. Countries with higher GDP may have better training programs, coaching, facilities, etc. which provide an advantage.
* Country life expectancy. The hypothesis is that countries with longer life expectancies have a healthier population and therefore have a higher potential to send better athletes to the compete resulting in a higher number of medals won.
* Country population. The hypothesis here is that more populous countries have a deeper pool or talent to build their Olympic team which results in a larger team competing in more competitions and increasing the potential to win more medals.
* Geolocation data of each country including country hemisphere and distance from the host city. The hypothesis is that factors such as travel distance, weather trends, or seasonality may have an affect on results. For example, the differences in seasons between the northern hemisphere and the southern hemisphere would mean the summer Olympics typically occur when the southern hemisphere is experiencing winter. Does this seasonality lead to underperformance by countries which are training in conditions atypical to summer Olympic conditions?
* Athlete gender. The hypothesis is that some countries may outperform in male or female individual or team competitions. A trend of outperformance in a certain discipline by male or female competitors may provide an advantage and lead to a higher expected medal count.

##Model Choice:

We utilized both supervised and unsupervised learning neural networks in our analysis. The supervised learning model was used to predict the number of medals a country would win at the Summer Olympics. The unsupervised learning model was used in our country and athlete demographic and geolocation analysis. The unsupervised model was used in the feature and demographic analysis because we were uncertain what insights and patterns we might find in the feature set and how they might affect Olympic competition results.

### Supervised Method

We used a supervised learning neural network model to analyze our results data. A supervised model was used because we have an explicit goal for our algorithm, predict the number of medals a country will win at the Summer Olympics.

### Unsupervised Method

This model was chosen because we were unsure what specific insights we might need to focus on when attempting to predict medal winners. Therefore, with a large amount of data collected and uncertainty regarding the specific features driving results we determined an unsupervised neural network would provide the most insight. Ideally, the model will help us group countries that would be expected to outperform and collect more medals at the 2021 Olympic games. Additionally, by analyzing feature importance we can construct additional predictive models which focus exclusively on the features with the highest influence on results.

Using the same thought process, by running the model on the individual results we hope to gain insight into whether specific countries outperform in individual events or disciplines. We also included gender as a factor to see if men or women from any countries are particularly dominant in their sport which would help boost the country’s medal count.

A limitation of our model choice

is the potential that the features we chose to analyze do not have a significant influence on competition results. Because we are using an unsupervised model it may be difficult to understand which features are important or if we are missing features which could be more influential to competition results. The nature of the unsupervised model makes it more difficult to tune and completely understand the input and output relationship.

The benefits of using an unsupervised model include the ability for the model to find patterns which are difficult to comprehend or analyze when simply reviewing the raw data. Because we are uncertain if our features do have a significant influence on Olympic results, an unsupervised model helps us identify important features and patterns