CSCI 4802 Project Proposition:
Olympics Data Analysis

Kin Seet
Sricharan Reddy Varra
Xingyu Zhou

February 2019

1 Our Previous Activities

1. **Kin Seet**
   Took Data Mining in Fall 2018 under Professor Qin Lv and done one project with group of three. We applied some classification methods to analyze crimes in Los Angeles such as clustering, linear regression and Chi-squared.

2. **Sricharan Reddy Varra**
   Last Fall I did Phase and Local Hack Day. I have taken applied probability last semester so I have a reasonable probability background, and currently I am in statistics with Xingyu learning R. However I haven't applied my probability and statistical knowledge, and that is what I wish to do in this project.

3. **Xingyu Zhou**
   I do not have so much knowledge about data analyze, but I am taking statistics class with R this semester, so it could help me a lot about how to use statistics knowledge to analyze the data set that we selected.

2 Project Summary

We will analyze the Olympic data from the 20^{th} century and the 21^{st} century so far and figure out various components of the Olympics that can be analyzed, and that will provide interesting results. This can include:

1. Rankings of countries for most medals in specific disciplines in the Olympics.

2. Analysis of countries in events over time.

3. Other geographic analysis on regions of the world that are excellent at certain events.

4. Summer and Winter Olympics and comparing and contrasting the results from both.
5. Most Improved Award in various events.
6. How population and GDP affects a country to win medals
7. Number of athletes in certain countries over time

3 Goals

The main goal of this project is to learn about the data science / analysis process and be able to apply it. Currently 2 of 3 of our team members are in a statistics course, and the other team member has taking data mining already so throughout the project we will get to practice methods of analyzing real data. With this general goal in mind, we can apply it to our Olympic data set, by figuring out interesting questions to find answers for and ways to analyze the data to acquire these meaningful answers.

Getting all of this data is interesting in itself, however designing creative methods of displaying the data and exploring it visually will be a goal as well. For instance a grid of various Olympic events could be made with countries flags on it for who ranked the highest in that specific event. Both R and Python have plenty of libraries to experiment and get very appealing graphics.

4 Techniques

4.1 Programs / Languages

1. R.
   (a) Using R for quick and dirty data analysis.
   (b) Use ggplot for data visualizations.
   (c) Map coloring is also available in R via ggplot
2. Python.
   (a) Numpy - A library that we can use to calculate the matrix, and for any math purpose usages
   (b) Scipy - A library based on Numpy, and also is scientific computing library
   (c) Pandas - A library based on Numpy as well, and use table to present data
   (d) Use the extensive data visualization tools that are available in the above libraries and others, also using python to do data cleaning to check if the data set has the valuable and reliable data.

3. SQL.

    (a) A database will the Olympics information

    (b) Extract subsets from the database and work on them in R and Python.

4.2 Methods

1. Clustering - Might use K-Means Clustering

2. Linear Regression - Population vs. Numbers of talented athletes. We would like to learn how population affects the number of talented athletes.

3. Time Series - Analyze the evolution of sports in countries. We would like to learn the improvement of sports from time to time.

4. Naive Bayes - With the result of time series, using Bayes's method to predict the performance.