Contents

1. Introduction

Our group work on the data analysis on dataset of "120 years of Olympic history: athletes and results". The process includes data collection, preprocessing, data analysis, modeling (Decision Tree and Log Regression) and GUI deployment. The goal of our project is predicting whether or not Olympians will win medal based on the historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. Out of the various features provided in data set, project will use some or all to build a model which will predict if a Olympian will win a medal or not.

The following is the workflow of our group. I do data acquisition and understanding. Sarvesh is in charge in Modeling and Jason focuses mainly on GUI Deployment. There is no absolute line between each step. As a team, we collaborate, work together and help each other whenever it needed.

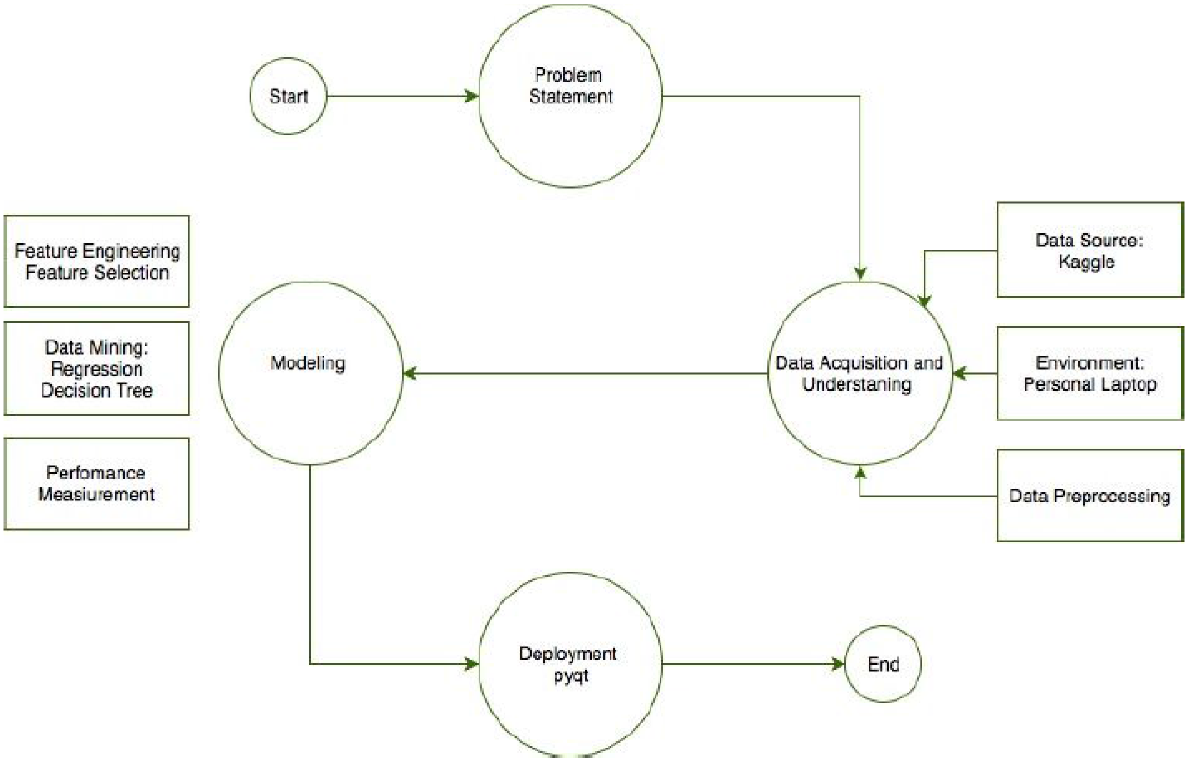

Start
Problem Statement
Feature Engineering Feature Selection
Data Mining: Regression Decision Tree
Perfomance Measiurement
Modeling
Data Acquisition and Understaning
Data Source: Kaggle
Environment: Personal Laptop
Data Preprocessing
Deployment pyqt
End

Figure 1: project workflow

I collected dataset of "Average body size in different countries" in order to fill up missing value of weight and height in preprocessing step. I did data analysis and visualization by Spyder in the perspective of sports, athletes, participation and received medals, which gives us an overview of dataset and a better preparation for modeling and deployment.

2. Description of individual work

Collecting dataset of “Average body size in different countries” in order to fill up missing value for athletes in preprocessing step.

Analyzing dataset in different aspects like sport, athletes, participation and received medals by Spyder.

Visualizing data by different methods like boxplot, barplot, pointplot, displot, heatmap and so on to get a better understanding dataset.

Working closely with other team member to write final report and prepare presentation slides.

Working as a good team member by setting up report template, scheduling team meeting, making reservation and so on.

3. Description of portion

Step1: Problem statement

My portion: Collecting and preparing dataset of "Average body size of different countries" for preprocessing.

Step 2: Data acquisition and understanding.

My portion: Taking main responsibility of data analysis on row dataset by getting the descriptive statistics and various figures/plots to get knowing the data; Using the results to prepare final report and slides.

Please see the code in my folder on Github and analysis result in the next 2 parts, result and conclusion.

Step 3: Modeling

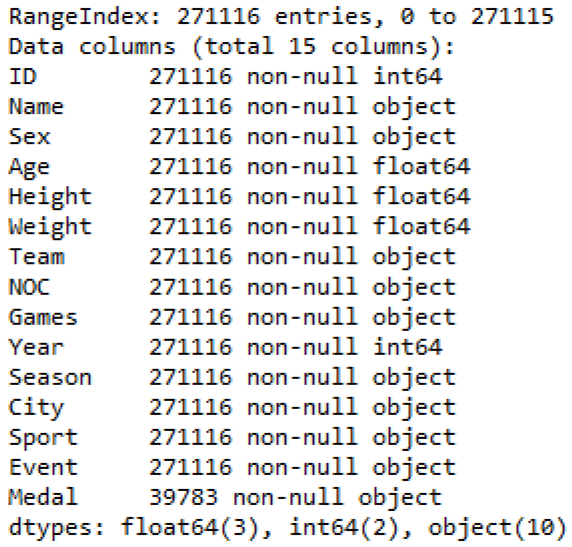
My portion: Testing the updated dataset updated from preprocessing by making use of the same code in step 2 to see how good the performance preprocessing does is.

Step 4: GUI Deployment

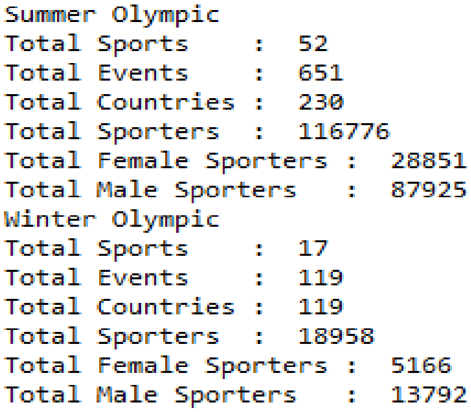
None.

4. Result

4.1 General Information



```
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
ID        271116 non-null int64
Name      271116 non-null object
Sex       271116 non-null object
Age       271116 non-null float64
Height    271116 non-null float64
Weight    271116 non-null float64
Team      271116 non-null object
NOC       271116 non-null object
Games     271116 non-null object
Year      271116 non-null int64
Season    271116 non-null object
City      271116 non-null object
Sport     271116 non-null object
Event     271116 non-null object
Medal     39783 non-null object
dtypes: float64(3), int64(2), object(10)
```



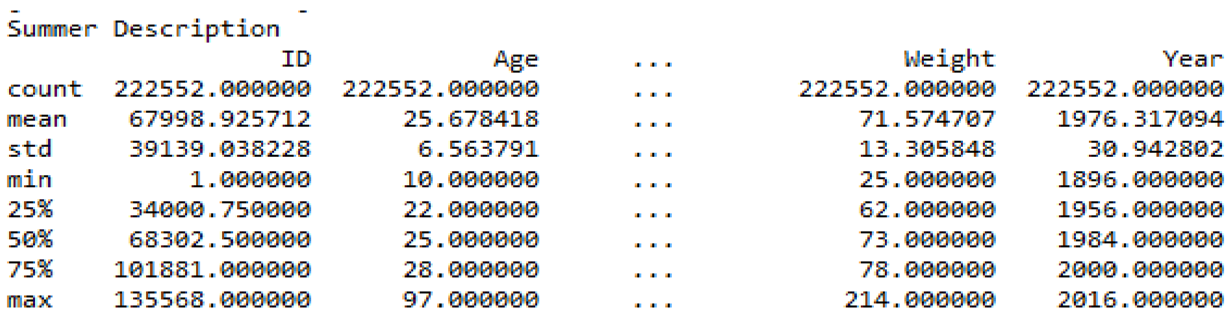
```
Summer Olympic
Total Sports     :  52
Total Events     :  651
Total Countries :  230
Total Sporters   :  116776
Total Female Sporters :  28851
Total Male Sporters    :  87925
Winter Olympic
Total Sports     :  17
Total Events     :  119
Total Countries :  119
Total Sporters   :  18958
Total Female Sporters :  5166
Total Male Sporters    :  13792
```

By using the code df1.info(), we get a detailed list of every column in dataset, including number of rows and data size.

Separating the total dataset into 2 subsets, summer and winter, we could do analysis on both subsets. There are more sports in Summer Olympic than in Winter Olympic. The athletes participated in Summer Olympic is ten times more than the one for Winter Olympic. For both games, the number of male athletes was more than 2 times than female.

4.2 Descriptive Statistics

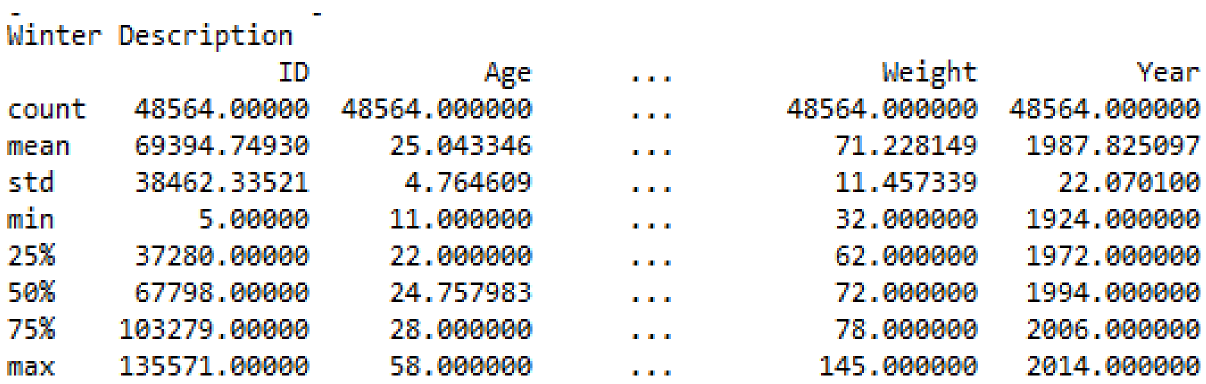
Summer



```
Summer Description
                  ID            Age     ...            Weight           Year
count  222552.000000  222552.000000     ...     222552.000000  222552.000000
mean    67998.925712      25.678418     ...         71.574707    1976.317094
std     39139.038228       6.563791     ...         13.305848      30.942802
min         1.000000      10.000000     ...         25.000000    1896.000000
25%     34000.750000      22.000000     ...         62.000000    1956.000000
50%     68302.500000      25.000000     ...         73.000000    1984.000000
75%    101881.000000      28.000000     ...         78.000000    2000.000000
max    135568.000000      97.000000     ...        214.000000    2016.000000
```

We get descriptive statistics table by code df1.describe. We have records as early as 1896, the oldest athletes was 97-year-old and the youngest athletes was only 10-year-old. A lot of interesting things could be found here.

Winter



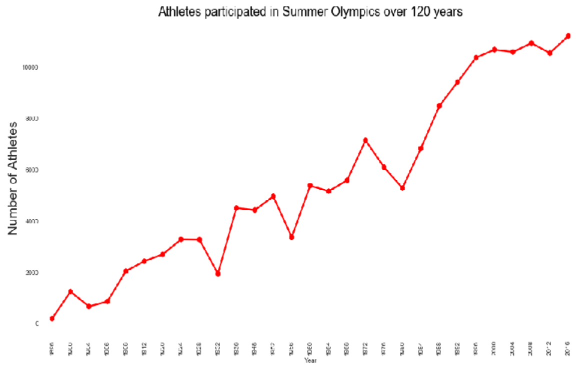
```
Winter Description
                ID           Age     ...           Weight          Year
count    48564.00000  48564.000000   ...     48564.000000  48564.000000
mean     69394.74930     25.043346   ...        71.228149   1987.825097
std      38462.33521      4.764609   ...        11.457339     22.070100
min          5.00000     11.000000   ...        32.000000   1924.000000
25%      37280.00000     22.000000   ...        62.000000   1972.000000
50%      67798.00000     24.757983   ...        72.000000   1994.000000
75%     103279.00000     28.000000   ...        78.000000   2006.000000
max     135571.00000     58.000000   ...       145.000000   2014.000000
```

Different from Summer Olympic, the oldest athletes ever participated in Winter Olympic was 58-year-old. And it looks like all the athletes were in great figures since the heaviest athletes was only 145 pounds.

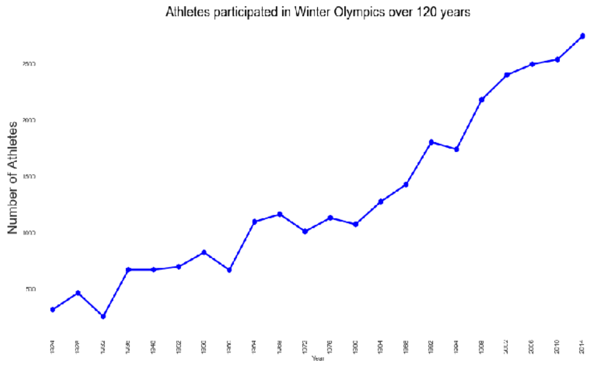
4.3 Participation Analysis

4.3.1 Participation by Athletes over Years

Summer



Winter



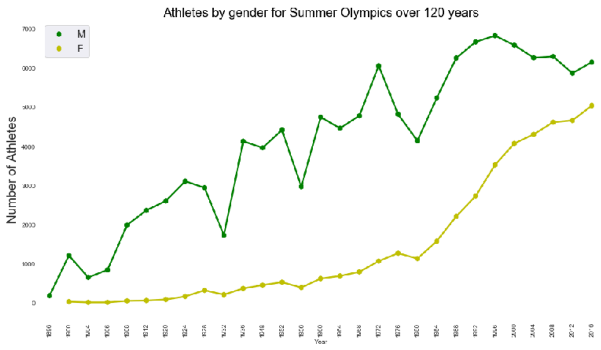
Over the past 120 years, the number of athletes participated in Summer Olympics has been growing fast. Because of the wars and boycott, there were some spikes in certain years.

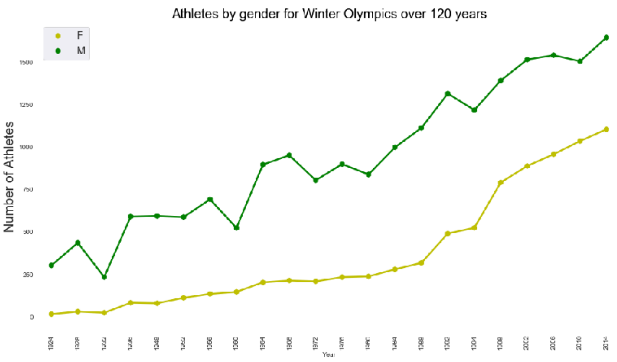
Compared to Summer Olympic, the number of athletes participated in Winter Olympics has been growing more smoothly.

4.3.2 Participation by gender over Years

Summer

Winter



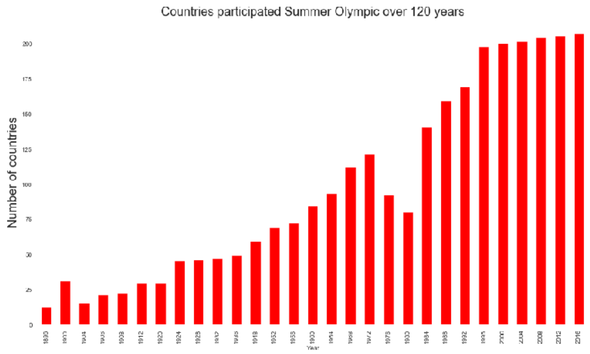


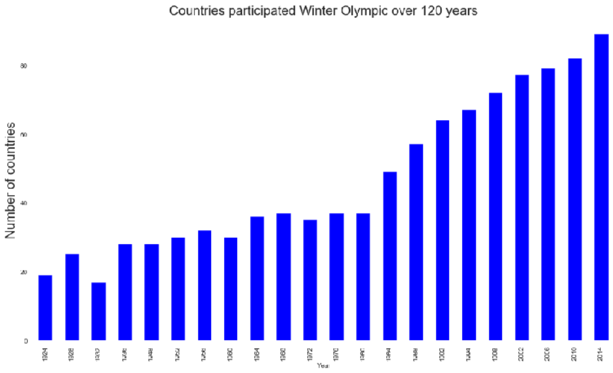
In every Olympic year, no matter it was Summer Olympic or Winter Olympic, the number of male athletes was always bigger than the number of females. Since year 1984, more and more female athletes participated in the Olympic games. It is nice to see that the gap between male and female athletes has getting smaller and smaller, especially in the past five Summer Olympic games.It is a great sign of civilization.

4.3.3 Participation by country over Years

Summer

Winter



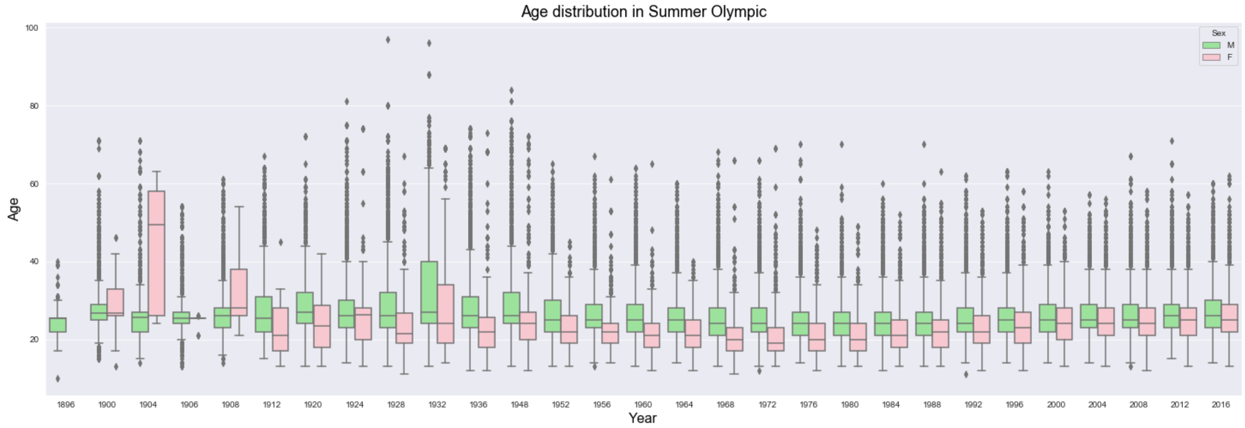


The number of participated countries is growing for both Summer and Winter Olympic Games over the past century. There was big decrease in both 1976 and 1980 because of the boycott.

4.4 Athletes Analysis

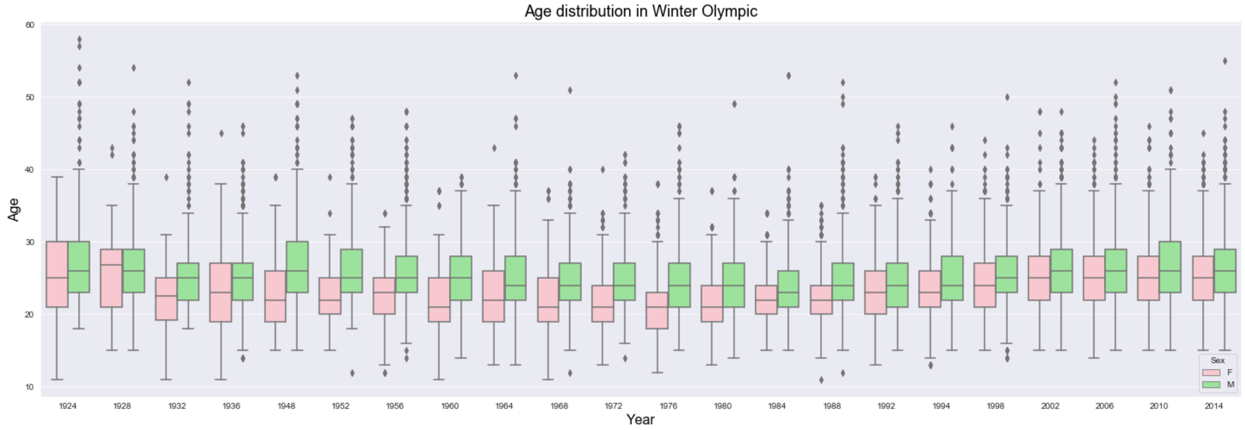
4.4.1 Age Distribution

Summer



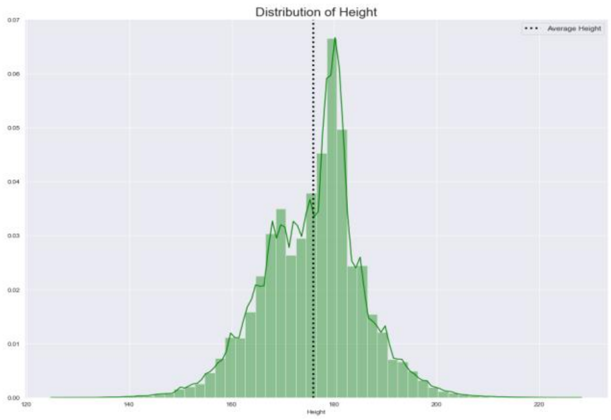
It is interesting that the age ranges were way wider in early last century for both male and female athletes. It is not hard imagine as the Olympic getting more and more popular, most countries would pick the best of best to participate, age limitation must be set for most of the situations.In every Olympic year, the male average age was older than the female average age.

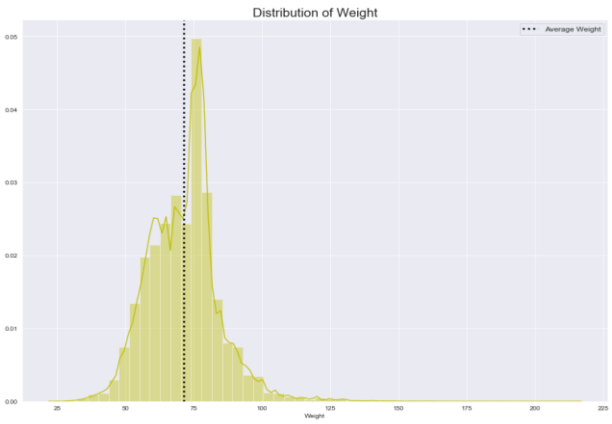
Winter



Compared to Summer Olympic, the age range in Winter Olympic is smaller in general. Just like summer Olympic, the male average age was older than female average age in every winter Olympic year. We see that for both Summer and Winter Games over time, the fluctuations in the distributions settle to being fairly constant. This provides evidence for dropping years that are below a threshold in order to reduce the noise in our dataset.

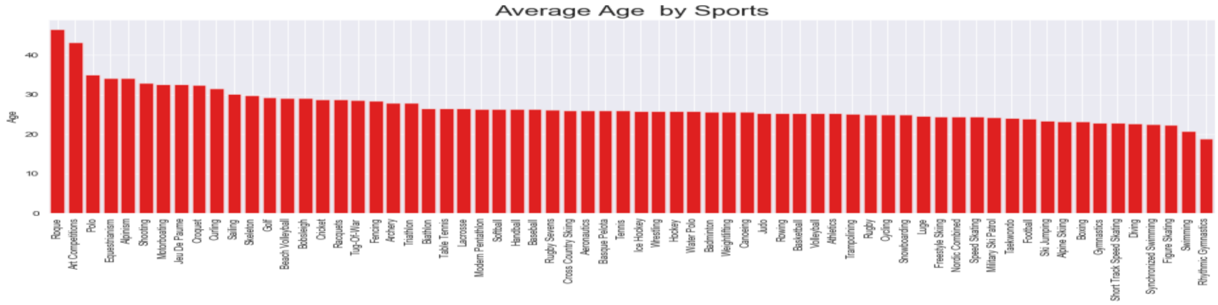
4.4.2 Height and Weight Distribution of Athletes

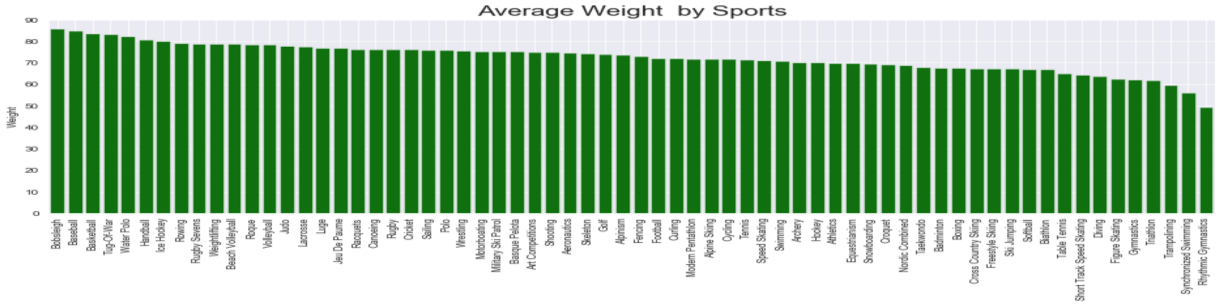


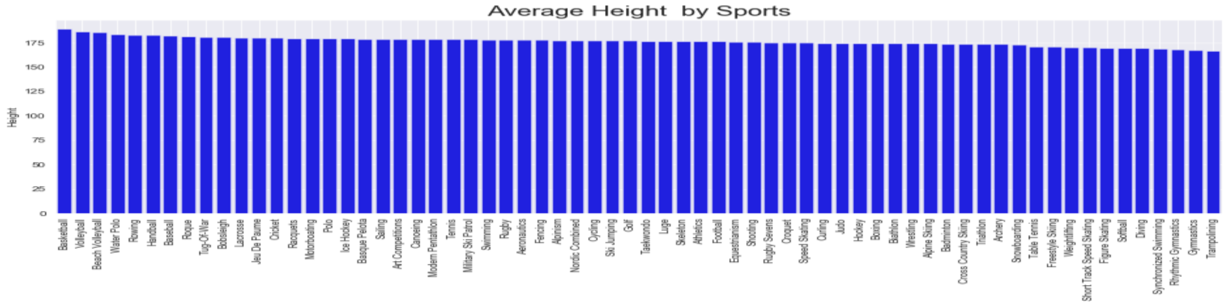


The average weight of all athletes is around 70 kilograms and the average height is around 176 centimeters. The distributions are nearly Gaussian, but are more skewed than we suspect. This may indicate that these variables will perform better in a non-parametric model that does not assume distributions for variables.

4.4.3 Average of Attributes by Sports





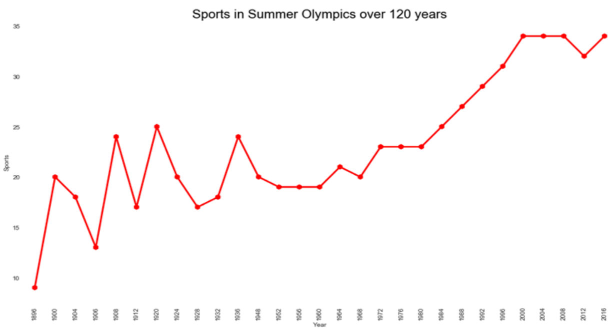


The average weight of athletes for sport Tug-Of-War is the heaviest. Like our common sense, the sport with highest average height is basketball. The average age of athletes for sports Art Competition and Equestrianism are way higher than others. This may indicate that a relationship between an athlete's physical characteristics and the Sport or Event will be important. This makes sense, as in general, taller athletes perform better in Basketball than shorter athletes simply be design of the sport.

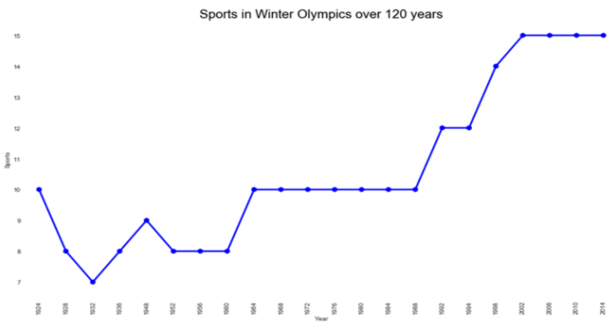
4.5 Sports Analysis

4.5.1 Sports over Years

Summer



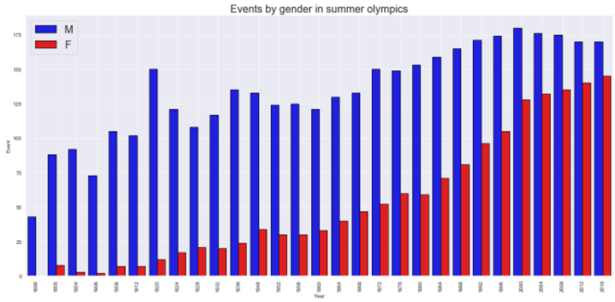
Winter



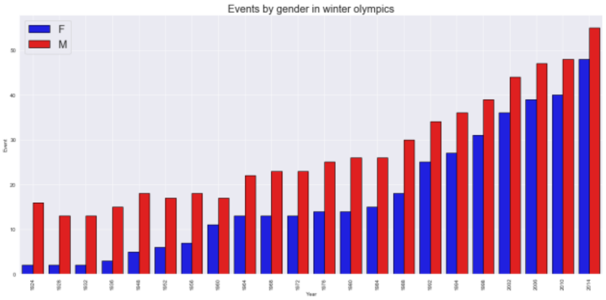
The events include all categories for Sport for both male and female athletes. More and more sports were added as Olympic Games. However, some Sports were also removed. This provides evidence that more recent data will be more important to our predictions, as old data may contain Events that are no longer played, or Events may have been played in the past that are not played anymore, such as Tug of War.

4.5.2 Events by gender over years

Summer



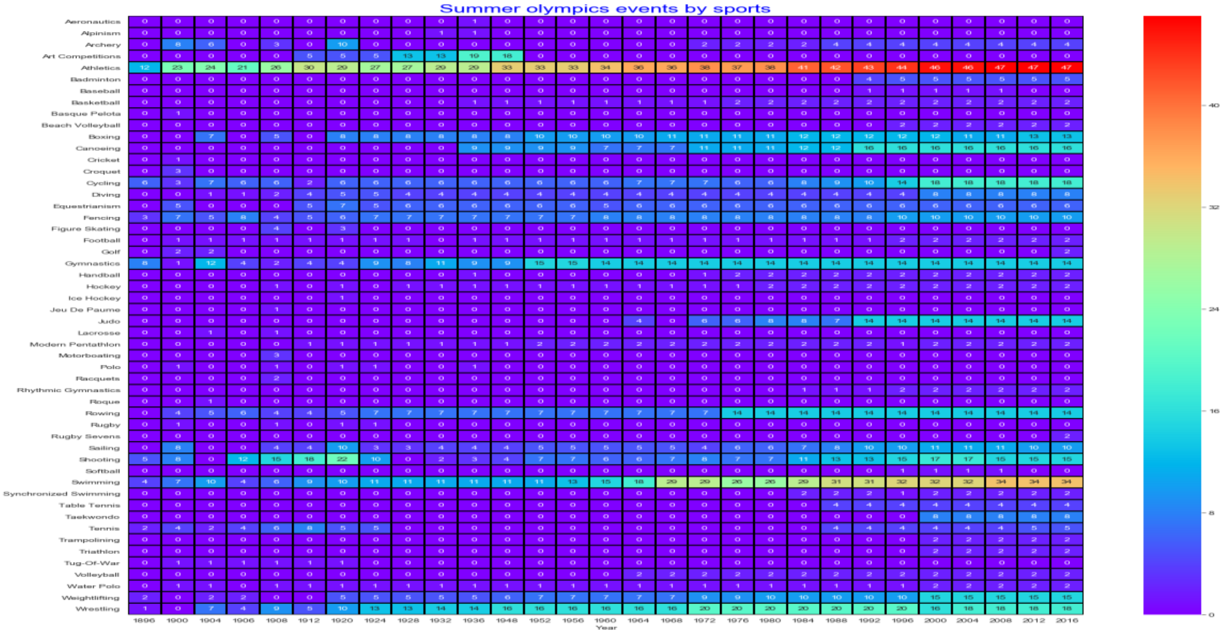
Winter



The events of women in the Olympics has been increasing since their first participation in 1900 for both Summer and Winter Olympics, and the gap between women events and man events has been getting smaller and smaller. Since Events are split by Sex, it may be that using more recent years will create more relevant predictions. For example, if we used only data with fewer amounts of women participants, the sample size for women may be too small to make a good prediction.

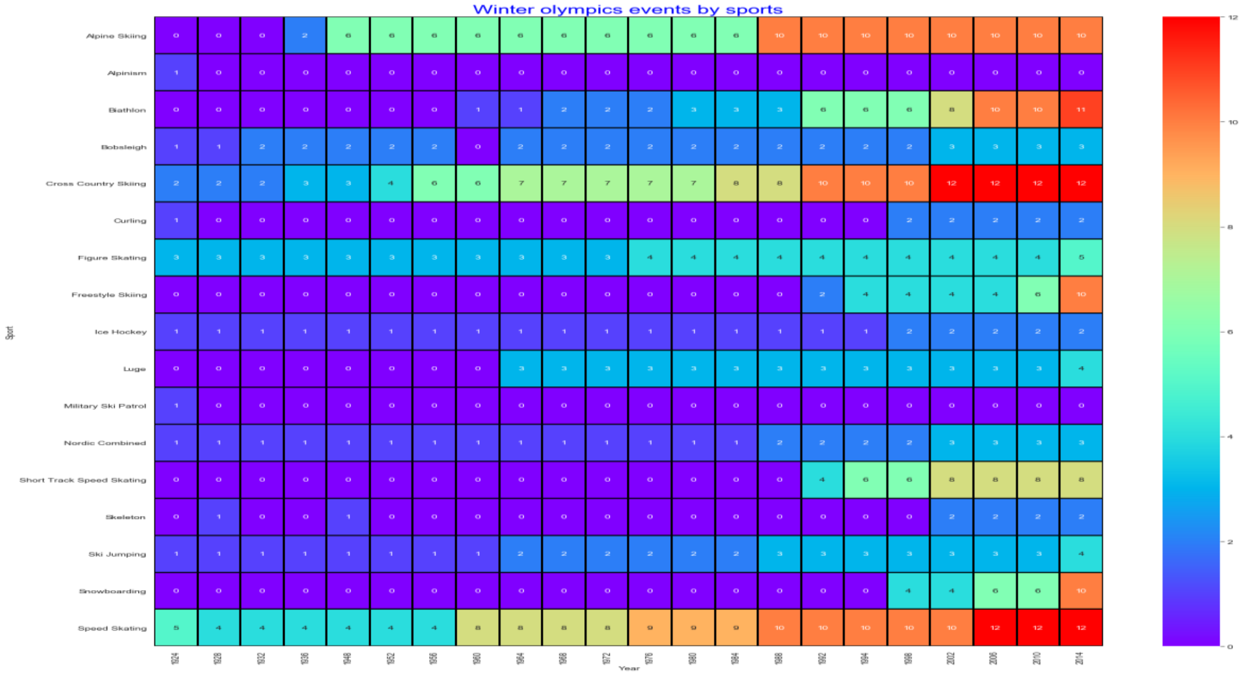
4.5.3 Event by sport

Summer



The numbers in each cell represents total number of events for every sport contested over the past years. The top 3 sports holding the most events in Summer Olympics are athletics, swimming and wrestling.

Winter

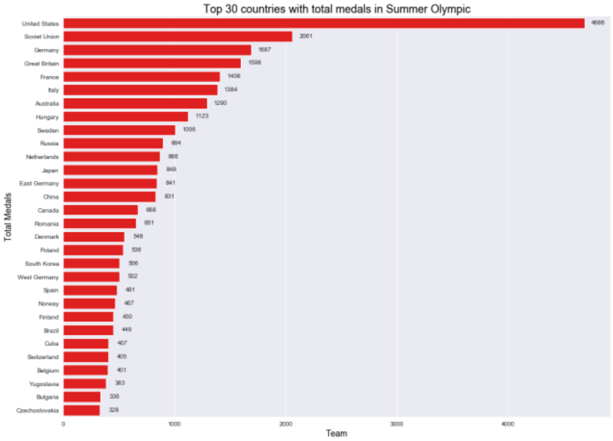


In Winter Olympic games, skating like speed skating, cross country skating, alpine skiing hold the most events in general. This indicates that the Sport will not be as important in predicting as the Event, as different sports contain different events. In order to reduce the noise in our model from this, we will remove the Sport column from analysis.

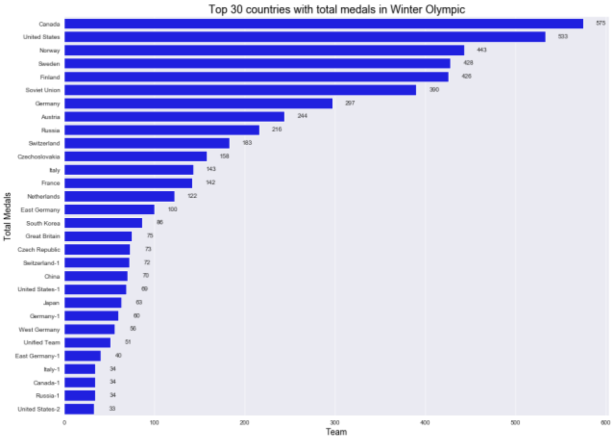
4.6 Medal Analysis

4.6.1 Top 30 countries with the maximum # of medals

Summer

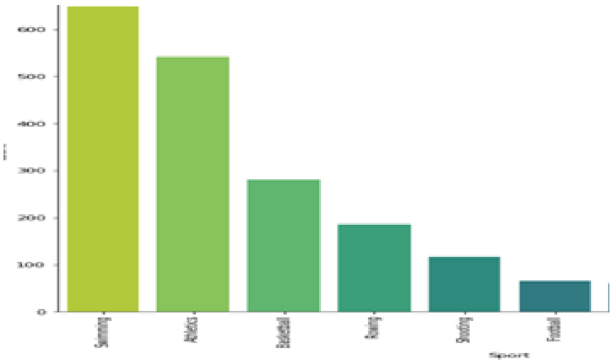


Winter



The top 3 countries with the most medals in Summer Olympic are United States, Soviet Union and Germany. The top 3 countries with the most medals in Winter Olympic are Canada, United States and Norway. From this analysis, we can conclude that the country of origin may be a good indicator of medal winners, as it seems that it is more likely that medal winners will come from the United States rather than Norway for example.

4.6.2 Medal by Sport for USA



The top 3 sports USA won maximum gold from are swimming, athletics and basketball in order.

5. Summary and Conclusion

The data exploration shows that a lot of data velocity exists in the original dataset. By dropping the records in certain years could improve our models. Combinations of variables that could be important, such as athlete body size (Height and Weight) and Event. It will likely improve our model by dropping redundant column like the Sports, as it seems to be similar with the Events. Country was also shown to be potentially important, as number of medal winners varies per country. It also may be that certain countries excel at specific events, such as the dominance of the Soviet Union in Hockey during the 1970's.

From the simple descriptive statistics table and all the plots, a lot of theory get confirmed. For example, male athletes in general have longer sport career and the number of participated male athletes is bigger than females. We get an overall understanding of the data set.

6. Percentage of code

p= (200-160)/(200+100)=0.12=12%

7. References

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#noc_regions.csv

https://wiki.python.org/moin/PyQt4

https://scikit-learn.org/stable/

https://www.worlddata.info/average-bodyheight.php

https://en.wikipedia.org/wiki/List_of_average_human_height_worldwide

http://www.wecare4eyes.com/averageemployeeheights.htm