**School of Information Sciences**

**(A Constituent Institute of Manipal University)**

****

**OLYMPICS: ANALYSIS OF SPORTS AND COUNTRY**

**SUBMITTED BY,**

**SHREE LAKSHMI G PRAKASH 171046014**

**N. SHIVA GANESH 171046026**

**COURSE:**

**ME-BIG DATA AND DATA ANALYTICS**

**GUIDED BY:**

**Dr B DINESH RAO,**

**ABSTRACT**

The modern Olympic Games or Olympics are leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate in a variety of competitions. The Olympic Games are considered the world's foremost sports competition with more than 200 nations participating. The Olympic Games are held every four years, with the Summer and Winter Games alternating by occurring every four years but two years apart.

* As a result of this numerous amount of data is being generated in every Olympics. And here we are considering Summer Olympics data for analysis in our project, and analysis following queries:
  1. Which Olympics had most medals won?
  2. Medals won in summer Olympics –By Countries.
  3. Top 10 Athletes.
  4. Which cities have hosted the games more than once?
  5. Male and female participation in each Olympic Games.
  6. Medals won in summer Olympics- By sports.
* USA team analysis.
* Predicting the time required to finish the 100m sprint(Men) in order to win Gold medal in upcoming Olympic game.
* Query which is used to find the trends in games, which takes years and a country as input and gives the corresponding details.

**1)INTRODUCTION**

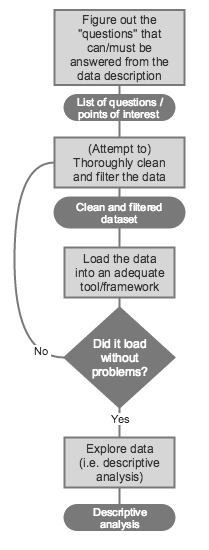
The modern Olympic Games or Olympics are leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate in a variety of competitions. The Olympic Games are considered the world's foremost sports competition with more than 200 nations participating. The Olympic Games are held every four years, with the Summer and Winter Games alternating by occurring every four years but two years apart.

Their creation was inspired by the ancient Olympic Games, which were held in Olympia, Greece, from the 8th century BC to the 4th century AD. Baron Pierre de Coubertin founded the International Olympic Committee (IOC) in 1894, leading to the first modern Games in Athens in 1896.

The evolution of the Olympic Movement during the 20th and 21st centuries has resulted in several changes to the Olympic Games.

The Games have grown so much that nearly every nation is now represented. This growth has created numerous challenges and controversies, including boycotts, doping, bribery, and a terrorist attack in 1972. Every two years the Olympics and its media exposure provide unknown athletes with the chance to attain national and sometimes international fame. The Games also constitute an opportunity for the host city and country to showcase themselves to the world.

**2)PROJECT FLOW:**

****

**Fig 2. Project Flow Chart**

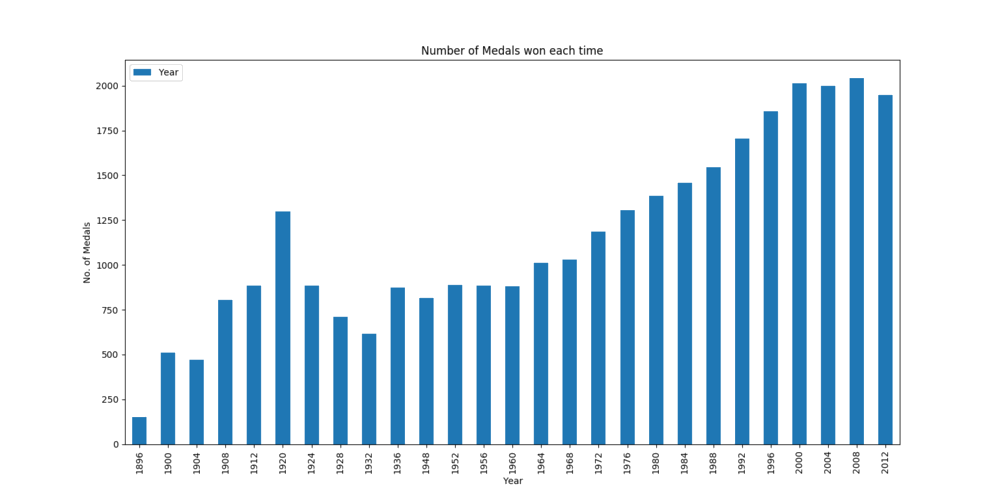
* After the first quick view, a more methodical approach must be adopted. The first step is to start asking questions that could potentially be answered by the data.
* The relevant data points that were previously identified must then be cleaned and filtered. The cleaning process can involve several strategies, such as removing spaces and nonprinting characters from text, convert dates, extract usable data from garbage fields and so on.
* The clean data can also be converted to a format (CSV, JSON, etc.) that will facilitate its loading into an adequate framework or tool.
* If some problem should occur during the loading process, it's likely that some detail escaped the cleaning process and consequently one or more of the previous steps should be reviewed.
* After loading the clean and filtered data successfully, the next step is to thoroughly explore the data. The main objective is to provide an insight into the data set, i.e., transform the raw data into actual informational content.
* The exploratory data analysis process involves things as the summarisation of the data, detection of outliers and anomalies or identify trends and patterns that could benefit from further study.

**3)DATA EXPLORATION:**

“Exploratory data analysis” is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there”

At this point, one must be able to answer the question "What do we want to do with this data?".

Now that we have framed the questions, let us see what our data needs to tell.

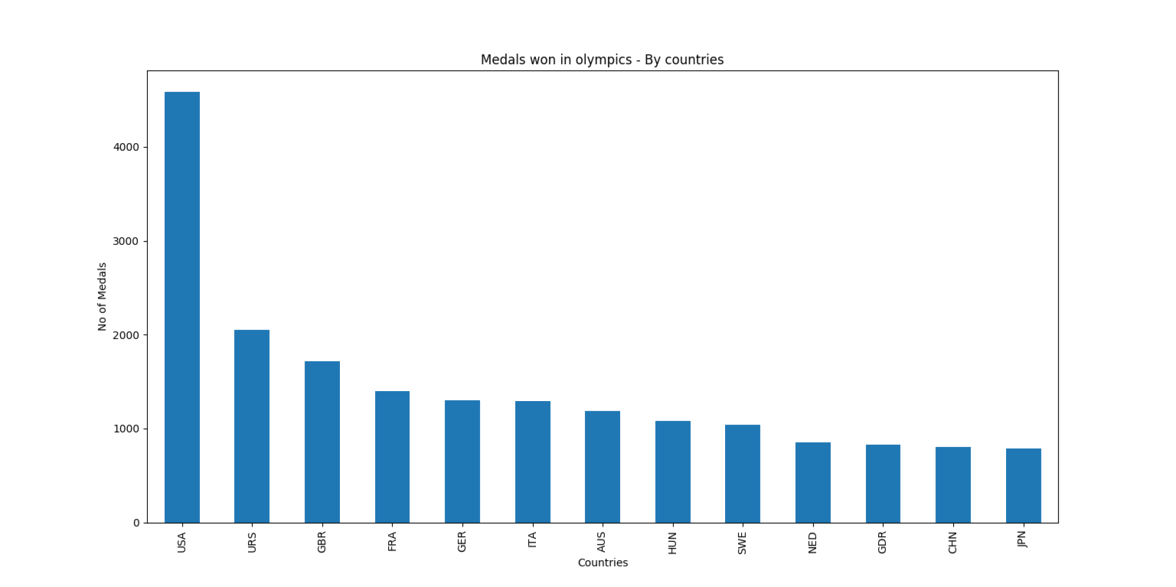
**3.1 Which Olympics had most medals won?**

**Fig 3.1 Number of medals won in each Olympics**

From this Visual, we can say that, number of medals won in each game is increasing as we move forward with the years. More predominantly from 1956 onwards. This can be due to increase in number of sports in each Olympics, increase in number of participating country and also due to increase in athletes in different games.

**3.2 What will affect the total number of medals won by different countries?**

Analyzing the number of Olympic medals won by geographic region by year reveals the true impact and extent of medal diversification. For example, whilst more countries are winning Olympic medals, how many medals are they capturing compared to traditionally strong Olympic nations? Is their success fairly minor or more pronounced? What are the possible contributing factors to their success in the Olympics?

Here is some of the reasons why some of the countries are winning more medals in Olympics.

**Fig 3.2.1 Number of Medals won by top 13 countries**

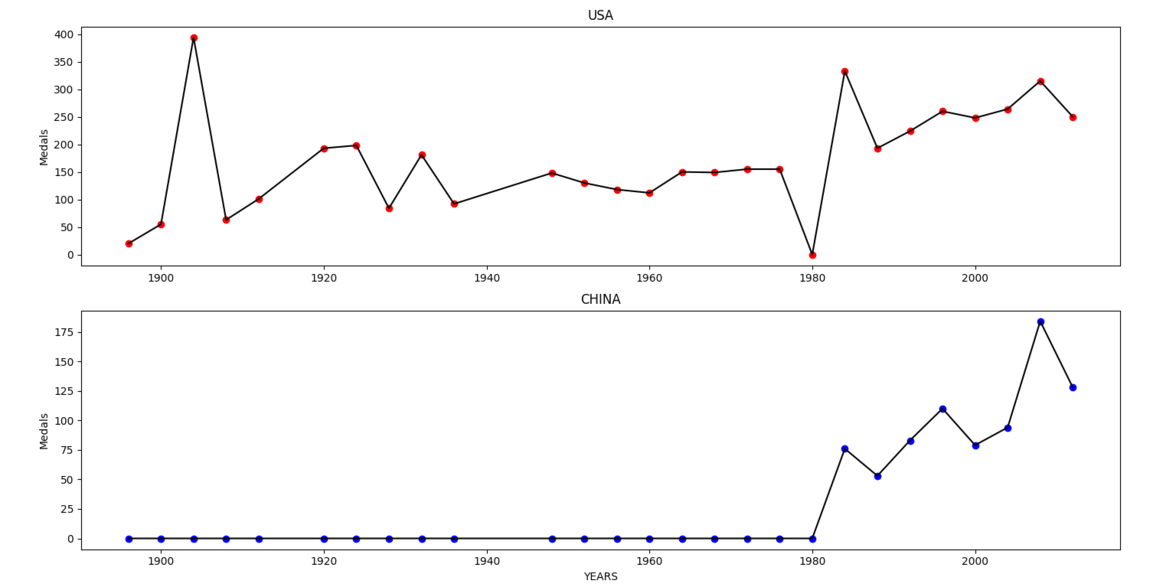
Top medal winning countries in entire Olympics is USA (United States of America), URS (Soviet Union) followed by GBR(Grate Britain) etc., And these countries are considered as richest and developed countries of the world.

Yes, developed countries win more medals than the developing or the underdeveloped. This is because in developed countries, people will be more aware of the things, they will get support from the government. The developed ones have a very good income. So, the countries can build coaching and training spots for different sports and can run them very well. They appoint well trained, physically strong for the sport, and qualified ones as coaches for the people.

Before World War II, Olympic success was dominated by the United States and Europe. Afterwards, more African and Asian countries begin to participate in the Olympics and the medal standings are marked by the arrival and growth of many regions including Japan, South Korea, China and Hungary.

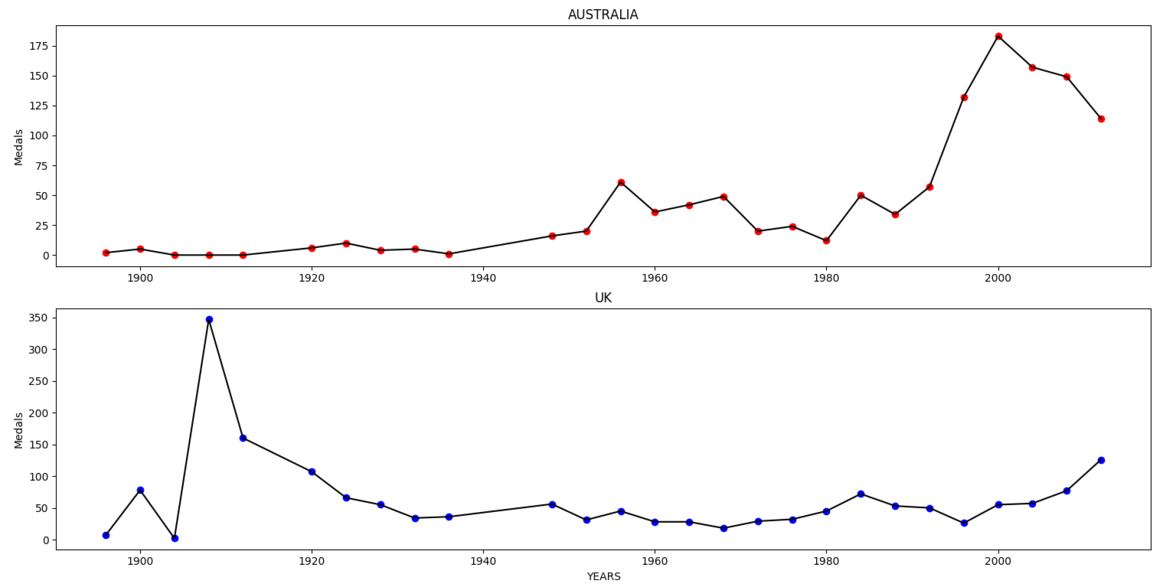
Here are three factors that may affect medal standings revealed by analysis:

* Past Olympics success.
* Host-Country Effect
* Future-Host Effect

1. **Past Olympics Success:** Medals won in the past can be seen as an indicator of “sports culture”.  The United States, for example, always perform quite well. Sporting prowess is important to them so, many people take part.
2. **Host-country effect:** The United States hosted the 1904 Olympics and won 394 medals compared to 55 at the previous games(see fig3.2.2 USA). The phenomena occur again and again. For instance, China hosted the 2008 Olympics and collected 100 medals compared to 63 at the previous Olympics(see fig 3.2.2 CHINA). This is a recognized pattern. Performing in front of a home crowd combined with extra investment in sport gives the host country a medals boost.

**Fig 3.2.2 Host-country effect (USA vs CHINA)**

1. **Future-Host effect:**

Australia won 27 medals in 1992 followed by 41 medals four years later (See fig 3.2.3 Australia). This was probably due to increased investment in sport in the run-up to the 2000 Sydney Games. The UK, as another example, increased its medal haul from 30 to 47 between 2004 and 2008, prior to hosting the 2012 Games (See fig 3.2.3 UK).

**Fig 3.2.3 Future-host effect**

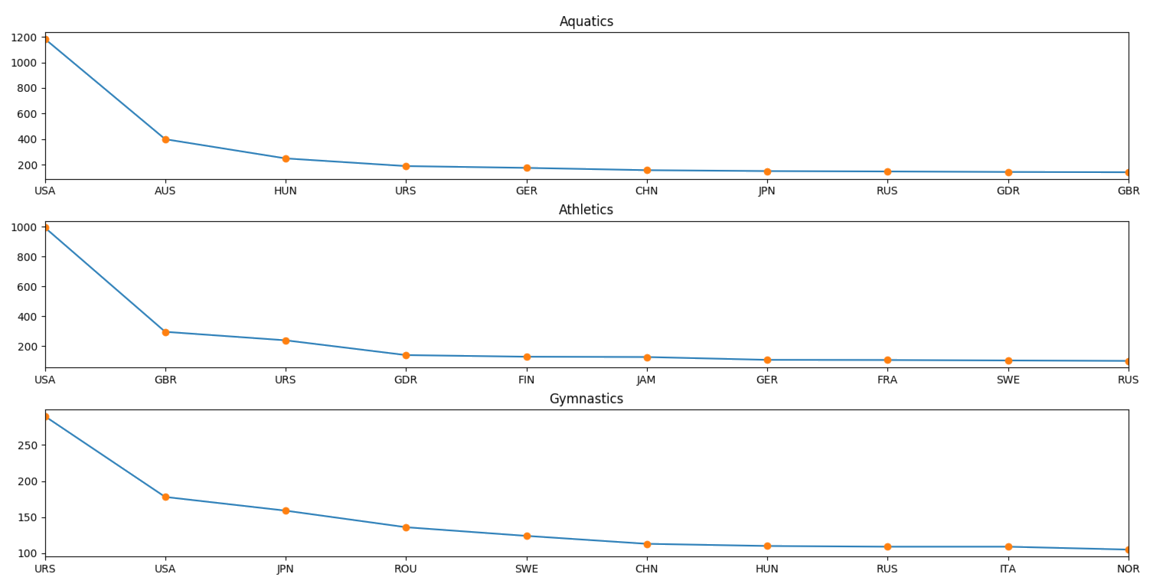
**Some other factors may be:**

**Wealth**: Countries with a high GDP, like Germany or the USA, can afford to invest in sports facilities and their populations have enough leisure time and money to take part in sports. This may not be the case in poorer countries.

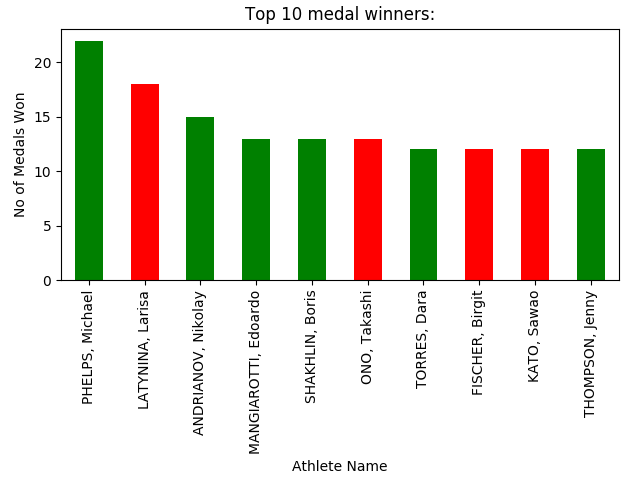
**Planned economies**: These countries with a high GDP growth tend to invest more in sport because they value the prestige that sporting success brings. China is a good example.

**Health**: Countries with a high life expectancy have a big healthy pool to choose athletes from such as Japan.

**3.3 Which country dominate 3 traditional sports?**

****Aquatics, Athletics and Gymnastics are considered as traditional sports of the Summer Olympics game. The three graphs (fig 3.1.1 ) show the United States, Australia excel in Aquatics, United States and UK are good in Athletics and Russia, United States and Japan dominates in Gymnastics.

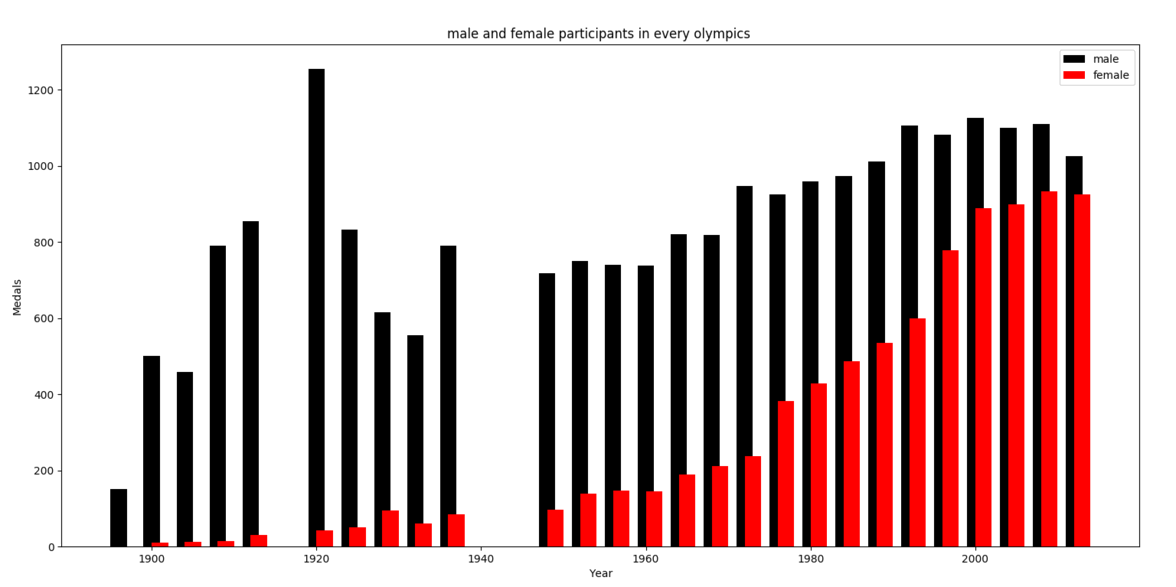
**Fig 3.3.1 Traditional sports analysis**

**3.4 Top 10 Athletes**

**Fig 3.4.1 Top 10 Athletes**

Michael Phelps, Larisa Latynina, Andrianov Nikolay, have been the most successful athletes over the entire history of Olympic Games. Michael Phelps is much ahead of every other athlete with 22 medals, a vast majority of his medals are gold medals.

In the fig (3.4.1), the green colored bara represent the male athlete where red represent the female.

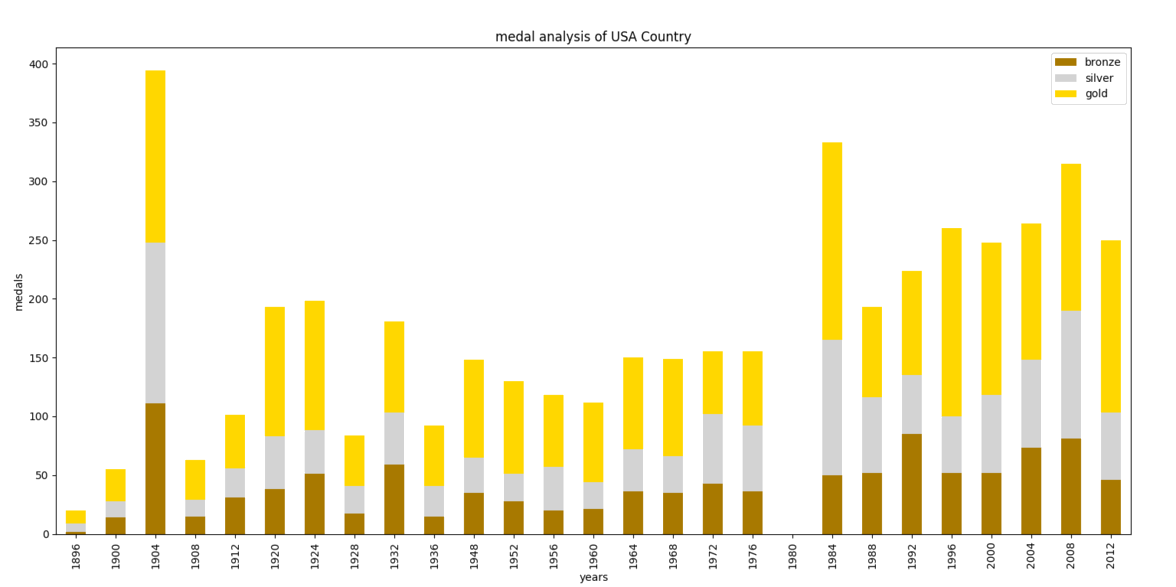
* 1. **Male and Female participation in Olympics.**

**Fig 3.5 Male and Female participation**

When the Olympic Games were first held in 1896, not a single woman had won a medal. The International Olympic Committee did not promote gender equality in sports for its first fifty years, but it gradually began changing its views on the topic and followed the lead of many progressive countries that were starting to view women as equal in almost all spheres of life. This attitude has led to a positive trend in which women are becoming more involved in sports, especially evident at the Olympic level. Not all countries have the same policies regarding women’s participation, but it can be said that women participation in sport has significantly increased over time.

In Figure (3.5) black represent the male participation in every year, whereas, red represent female participation.

**3.6 Team USA - Facts on funding and building the team**

**3.6.1 USA Country analysis**

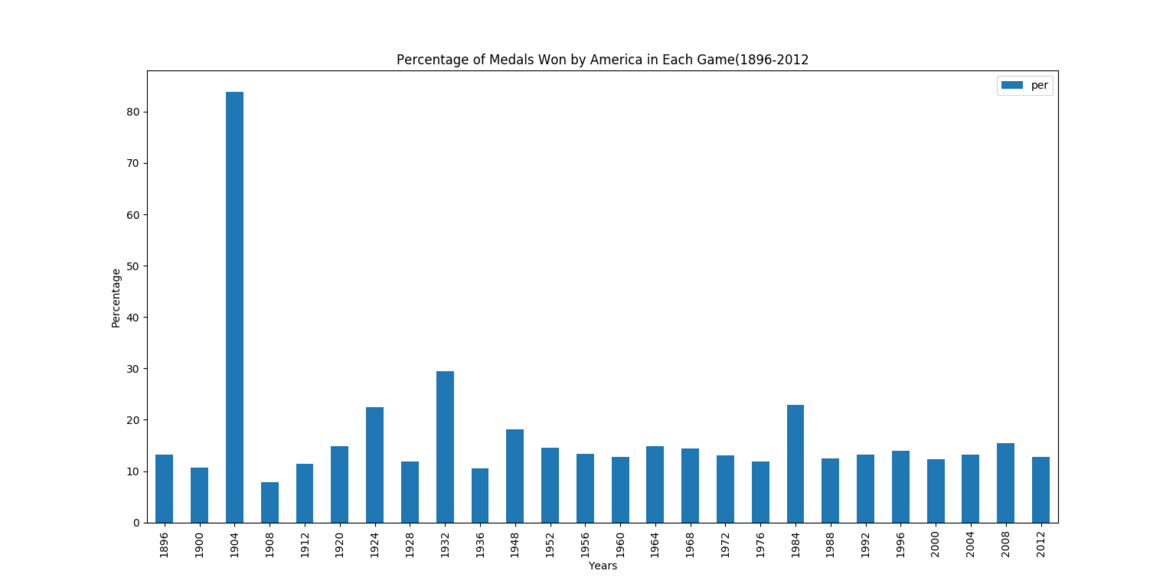
A close up of a logo

Description generated with high confidence

**Fig 3.6 Analysis of USA**

Figure (3.6) is a stacked graph used to show the number of Gold, Silver and Bronze medals won in each Olympics. Here Yellow represent number of gold medals, silver represent number of silver medals and brown represent number of bronze medals in each Olympics.

Interestingly the best game for USA was in 1904. Second best is 1984 which looks a bit suspicious as there are no entries for 1980. American Athletes love GOLD. Nearly 50% of all the medals won by Americans are Gold. (Shown by Pie Chart inside Fig 3.6)

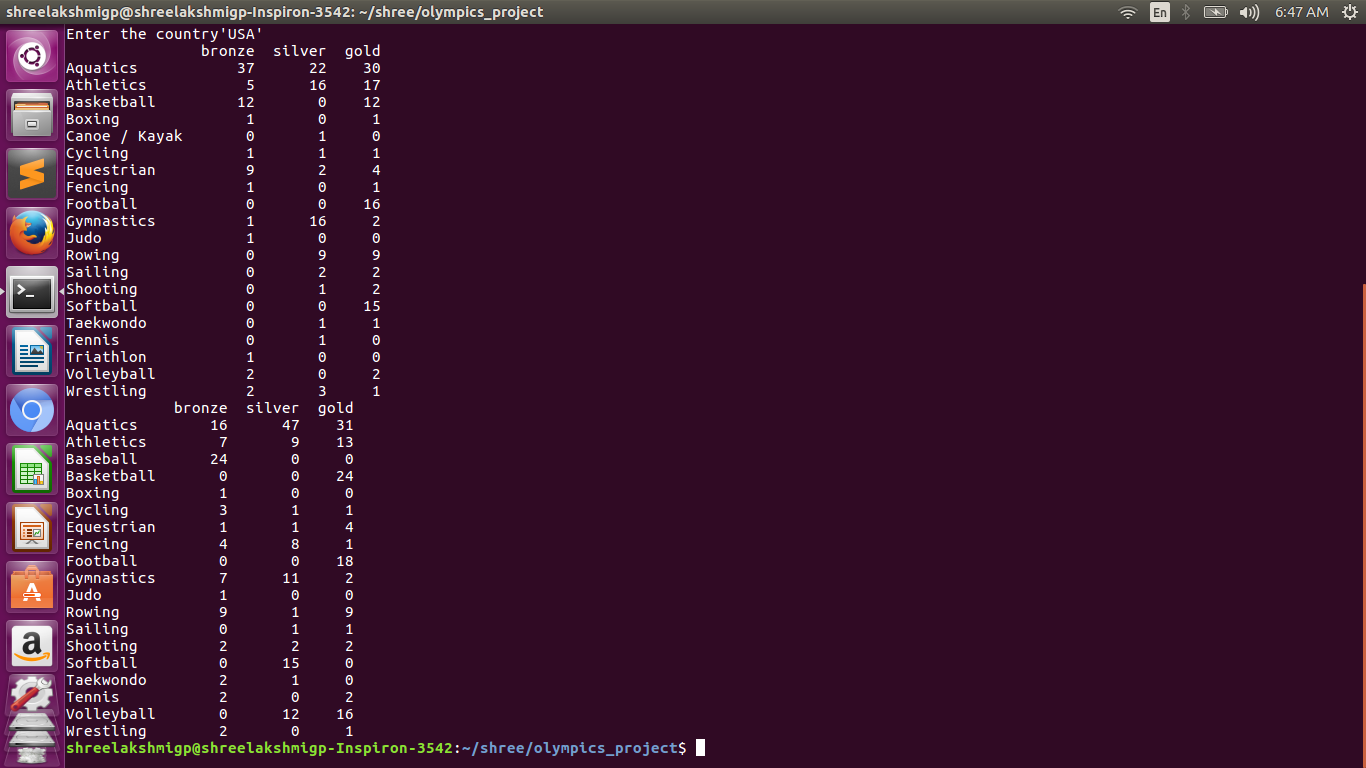
**3.6.2 USA-Percentage of medals won each time.**

**Fig 3.6.2 medal analysis of USA**

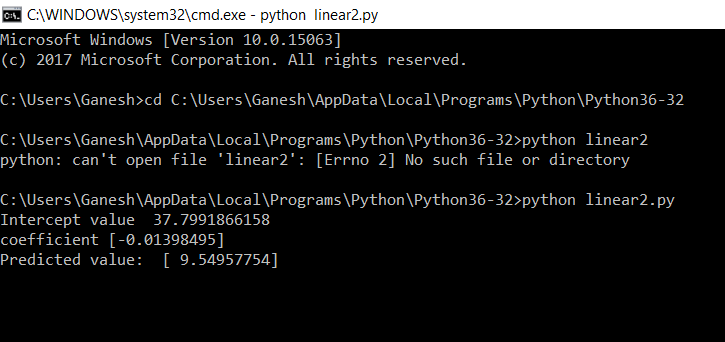
1904 Olympics was hosted by USA. More than 80% of them were won by USA. America is consistently winning more than 12 percent of total medals in all the games.

**4)Modelling:**

**Query:** To find the trends in games, which takes years and a country as input and gives the corresponding details.

****

**4.1) Linear Regression**

* Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.
* A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).(Figure 4.1.2)
* Objective of using Linear Regression is to predict the time required to finish the 100m sprint(Men) to win Gold medal in upcoming Olympic game.
* We collected the data on 100m sprint Men into a csv file and applied linear regression by taking Year and Time (seconds) as parameters for which the medal equal to Gold.
* The Output(Figure 4.1.1) gives the prediction of Time(seconds) in which an athlete need to finish to win a gold medal in the future Olympic game.

**Fig 4.1.1**: Values of Linear regression

A screenshot of a cell phone

Description generated with very high confidence

**Fig 4.1.2:** The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

**5)REQUIREMENTS:**

**HARDWARE REQUIREMENTS:**

* 40GB Hard Disk or above
* 2.0GHZ clocked speed dual core processor or above
* 4GB RAM or above
* 2GB Graphic Card

**SOFTWARE REQUIREMENTS**

* Tool: Python 3.5.3 or above, Microsoft Office, Sublime editor, PyCharm
* Libraries: Numpy, Pandas, Matplotlib, Sklearn, SciPy

**6)FUTURE WORK**

In future, we would like to apply K-means clustering and investigate whether the same pattern exists, and, if there is difference, what is the reason for this difference.

We applied Linear regression for only one event, 100m sprint Men and predicted the time required to finish the race in future Olympic game. In future, we can apply for more events and predict the results for future.

**7)References:**

* <https://www.olympic.org/ancient-olympic-games/history>
* <http://www.history.com/topics/olympic-games>
* <https://www.bloomberg.com/graphics/2016-olympics-usoc-return-on-investment/>
* <http://www.teamusa.org/us-olympic-and-paralympic-foundation/team-usa-fund>
* <https://en.wikipedia.org/wiki/1904_Summer_Olympics>