Final Report

Group 1

Zach Beisler
Billy Gault
Morgan Sterling
Zack Vliet

Introduction

When approaching this project as a group, we wanted to create a webpage visualization that was both informative and allowed for user interaction. Data visualization seems to be most effective when the contents are all displayed on one page, and the user can interact with the webpage in order to analyze the data how they would like. This is a very involved process, and we knew that there would be a lot of work in order to create a dynamic visualization tool for a specific data set. We were unsure what problem we wanted to address, so we decided to look for a data set before formulating our problem. Upon searching, we found data on all of the Olympic games from Athens 1896 to Rio 2016. While the data was interesting, we quickly learned that it was challenging to draw any conclusions from just the raw data. We decided to solve the problem of visualizing what countries won the most medals during a certain year. We also wanted to visualize the amount of athletes representing each country.

Our goals for design were to have some kind of map displaying the countries of the world. This map would be used to gain insight on medals and athletes. We were unsure how, exactly, we wanted to allow user interaction, but we decided dropdown menus would be applicable. Our design would require the use of D3.js, php, and a map API to display the data. We would also utilize a MySQL database to hold the data and query from it. We had many ideas for the kind of map we wanted to use, but decided on a simple map of the world after many iterations and a lot of discussion.

Data and Preprocessing

The dataset we used is on Kaggle.com and it is a history of Olympic medal winners from 1896 to 2016. There are 271,116 data points, and they are described by 15 features. The features include ID number, name of the athlete, sex of the athlete, age, height, weight, team,

NOC (National Olympic Committee 3 letter code), the Olympic Games, year, season, city, sport, event, and medal. There is also an accompanying table that specifies the region of each NOC if a country's code has changed over time. Of the roughly 200,000 data points, not each one is a unique athlete. There are repetitions in athletes as some athletes compete in multiple events in a single Olympics.

The preprocessing was very important because, with this data and its features, there are many different ways to visualize. However, being that we were only interested in a couple aspects of the data we were able to eliminate much of what was provided. Specifically, we were interested in the amount of medals, amount of athletes, type of medal, and the countries that they correspond to. We began by downloading the data and pulling it up in Excel and R to take a look at it. This step is very important because humans have better judgment than computers, and getting a person to look at the data first will provide insight into what is missing, how the data is laid out, and and how it is represented among other things. Upon inspection, we were able to eliminate many of the unnecessary feature columns to leave us with only ID number, athlete, team, NOC (National Olympic Committee 3 letter code), year, and medal. We went through the process of creating an acceptable MySQL database and stored all the data as a simple flat file. This eliminates the need for joins and increases overall speed.

Visualization Design

We originally thought a heat map would be the best way to display this data. Over time, the 'hot' areas would shift and the user would be able to observe gradual changes on the map. However, we learned that heat maps are very general. They focus on general area rather than specific countries, so it would be nearly impossible for the user to determine what actual country had the most medals and athletes, which was contrary to the problem we set out to solve.

Therefore, we decided to use a simple world map to show distribution of medals. A slider would be used to select the year of Olympic games, and the color of the shaded country would correspond to the amount of medals that country won (i.e. darker the color = more medals). We decided on a slider rather than a drop down menu because of the ability to see change over time. Drop down menus are abrupt and less helpful since it takes more time to switch between years and it is difficult to follow these shifts. Additionally, in order to give the user specific information, we added a hover feature to the visualization. Once the mouse is hovered over a country, the actual medal count will appear for that year, giving the user even more insight to the specifics behind the map. For additional user interaction, we provided a dropdown menu to select a color scheme preference (blue, orange, or green). Then, as the final aspect of visualization, we provided a 'Play' button enabling automation of the slider. This allowed one click and the slider would move according to time, showing trends in olympic medal distribution effectively outlining historical events that could have affected the Olympics (World War I, World War II, etc.).

Following our presentation we wanted to add one more aspect to the visualization. We decided that we wanted to incorporate a bit more analytics, so we added a scatter plot enabling the user to select and compare countries over the years. This gives the user more insight on the numbers behind the map and it allows the user to see trends over time.

Implementation

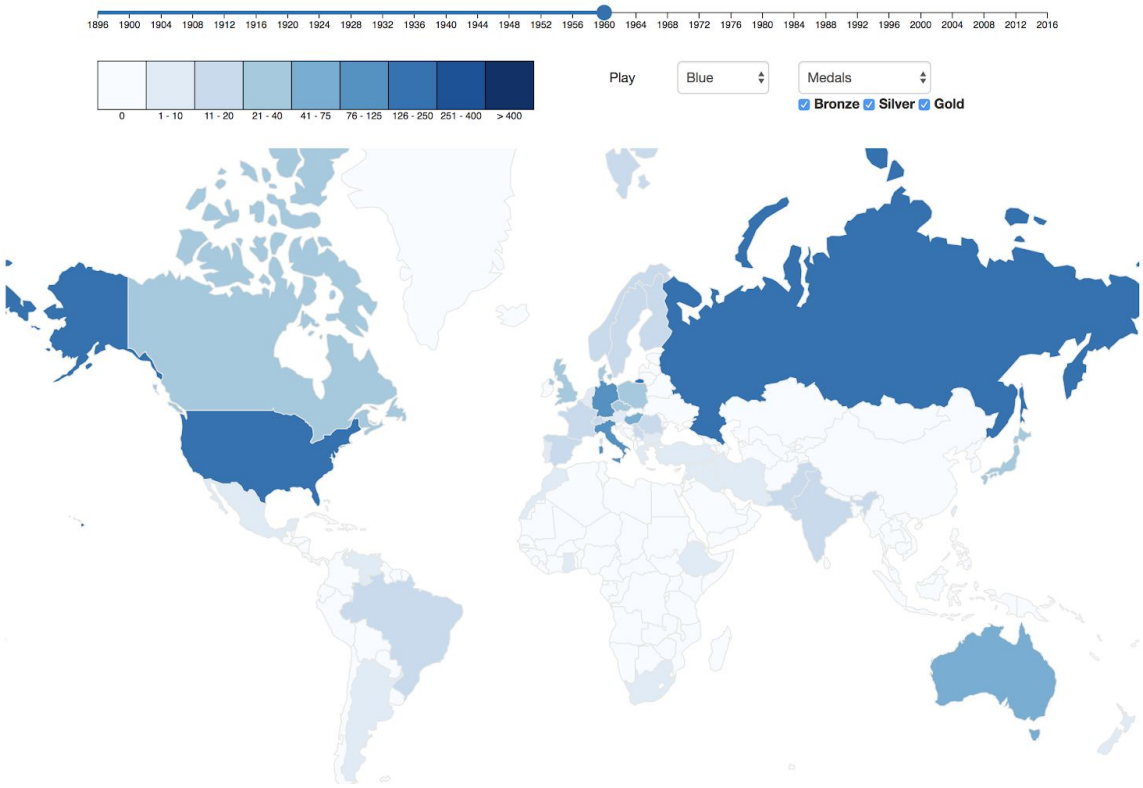
As mentioned before, the first step in the process is, of course, to analyze and process the data. We chose to do this in Excel and R since we have background in the language, and it is a very powerful when it comes to data processing. Processing the data in R allowed us to see the data more clearly. Upon analysis we were then able to determine exactly what our dataset

contained. Talking as a group, we made a decision on how we wanted to filter and visualize our data. After this decision, we were able to move forward, and thus needed a place to store our data for future access. We chose a simple way to store our data with PSU's database management system, phpMyAdmin. Our site is hosted on the PSU webspace where we support an animated visualization system that counts medals, events, and athletes per olympic region over time. This animation is supported by jQuery, D3js, and datamap. Due to the nature of the region encoding in our data, we had to come up with a lookup table to map the regions in our database to the country codes supported by the Datamap API. For example, China is mapped to CHN. However, the USA is represented by "United States of America" not "USA," which was not supported by Datamap. Therefore we needed to write a small script to fix the mappings for a handful of countries and regions.

Results

We are very pleased with the results of our visualization. It is dynamic and provides lots of user interaction for the best possible user analysis. Figure 1 shows an overview of the entire webpage. As you can see, we have the world map layout with the slider above and the multiple drop down menus next to it. The color scale is used to show a glimpse of the medals won by each country and athletes participating, based on color.

Figure 1:



Figures 2, 3, and 4 show a closer look at the slider as well as the drop down menus and their options.

Figure 2:

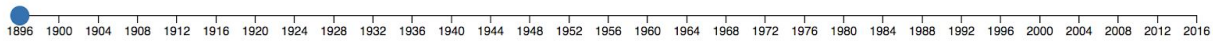


Figure 3:

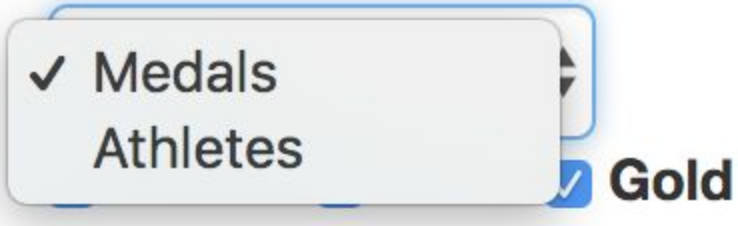


Figure 4:

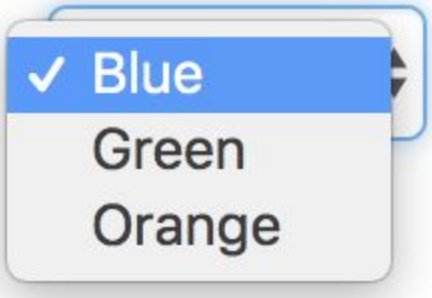
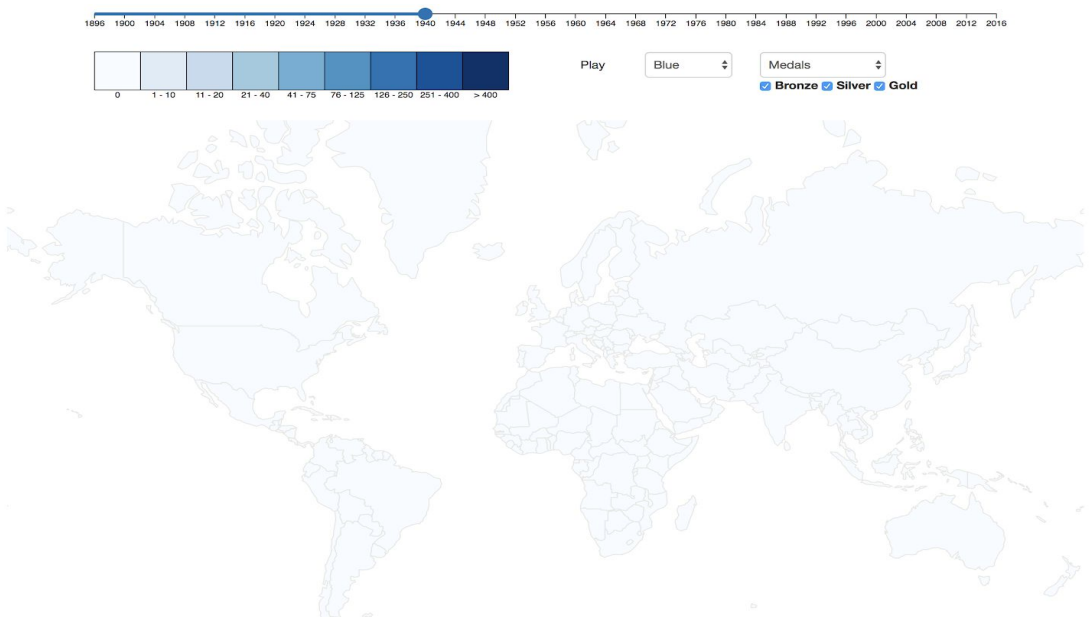


Figure 3 shows the dropdown which allows the user to choose whether they want to see the amount of medals won by each country, or the amount of athletes that each country has participating in that specific Olympic games. Figure 4, shows the menu allowing the user to select between the color schemes of the map. We also found some interesting information from the map. For example, there were no Olympics during 1940 and 1944 due to World War II, therefore there was no data for these years. This can be seen in Figure 5 below.

Figure 5:



Figures 6 and 7 display the map in the two other color schemes, green and orange. Additionally, Figure 6 displays no medals won by any country during World War I, and figure 7 shows that, following WWII, Germany and Japan were not invited to participate, while Russia, then known as the Soviet Union, chose not to participate at all.

Figure 6:

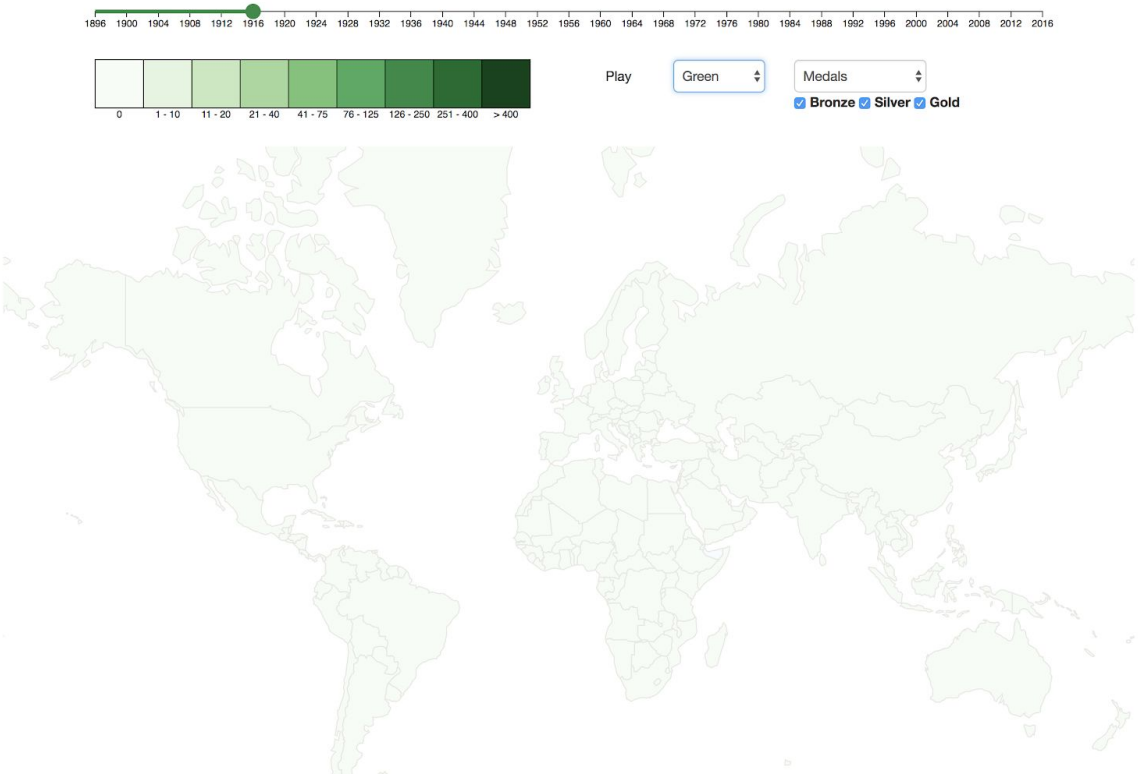


Figure 7:

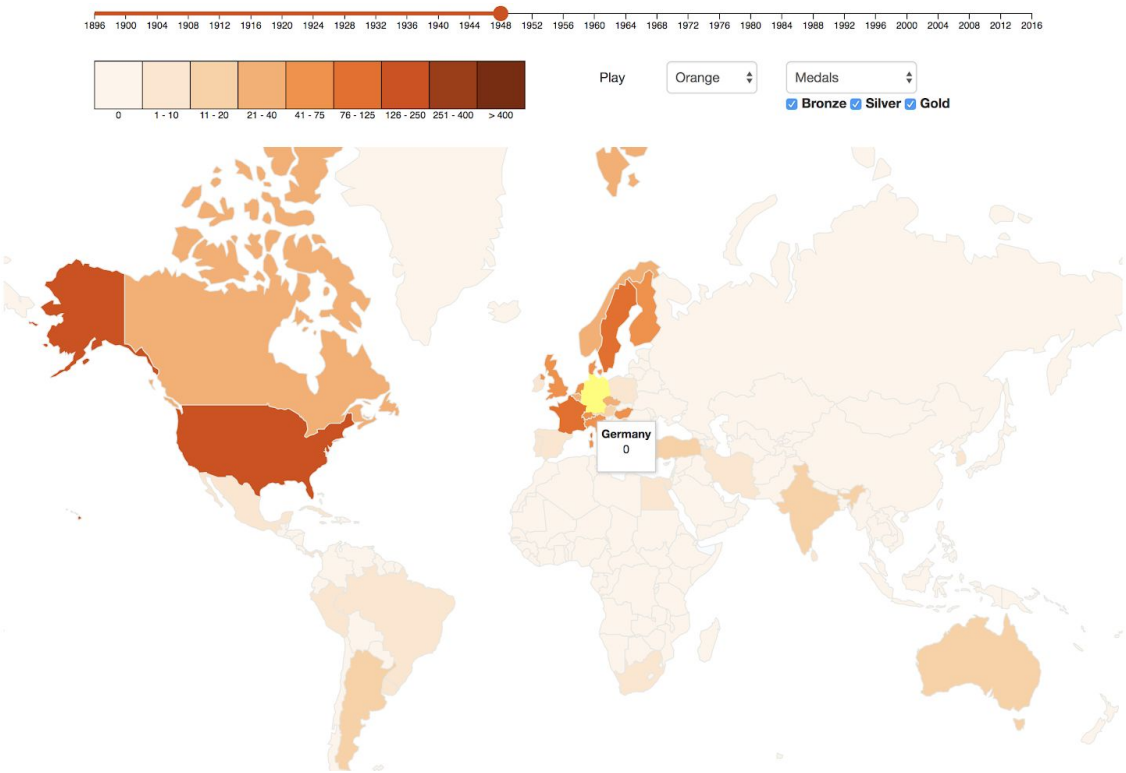


Figure 8 shows the menu at the top of the screen allowing the user to choose between map and scatterplot. Meanwhile figure 9 is a picture of the scatterplot comparing the USA and Russia medal counts over the years. You can see where one country wins significantly more medals than the other, and this is 1980 where Russia clearly had the upper hand. However, in historical context, this was the year that the United States boycotted the olympics due to the Russian invasion of Afghanistan, explaining the difference in medal counts.

Figure 8:

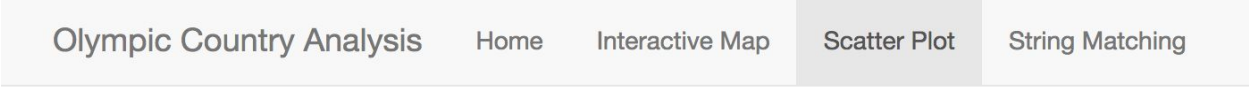
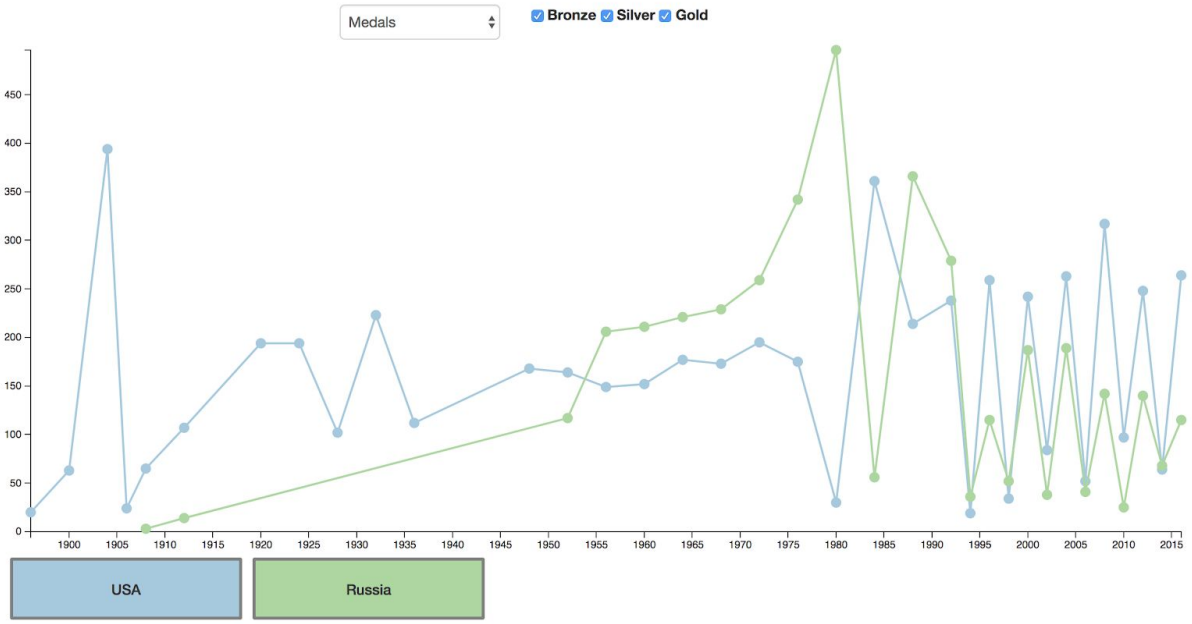


Figure 9:



Conclusion:

Our group learned many valuable lessons by completing this project. For one, we learned many skills, such as D3.js, that are incredibly powerful in data visualization. It is a

challenging language to learn, but, once it is practiced enough, it is one of the most powerful data visualization tools available. We also learned that we cannot try to accomplish too much in too little time. We originally intended to have histograms included on our webpage to support the map. This would give the user information on medal distribution among other things. However, the map consumed too much of our time and we were unable to include these graphics. Although they were not essential to the user experience, they would have been very informative. We did, however, manage to complete the scatterplot following our presentation, which is very beneficial to the user. We also were able to brush up on our SQL and php in the code and gain some more skill in those areas. As far as team collaboration goes, our group worked very well together. Our more design oriented teammates were essential to planning the layout while the more technical teammates worked well together in writing the code and debugging, although everyone worked on all aspects in some way. If we were to change anything we may have more group meetings earlier in the process, but overall the process went very well and the final product shows our hard work and collaboration efforts.

Appendix

Url: https://my.up.ist.psu.edu/zqb5073/map.php

Source Code: There was too much to add here so we are submitting the source code as a file along with this report.