DATA ENGINEERING 2: DIFFERENT SHAPES OF DATA

Department of Economics and Business

Central European University

2020/21 Fall

2008 Summer Olympics: A Use Case Study on Data Product

In this project, we study the correlation between the number of Summer Olympic medals each country won in 2008 and their GDP per capita, population, and sports expenditures using different data tools and concepts. We find that medal count and population are positively correlated, same with GDP per capita and government expenditure, but the later two are not statistically significant at 10%.

Team 8

Abduvosid Malikov 2002824

Li Deborah Jia 2000692

Xinqi Wang 2003316

1 Introduction

Due to the COVID-19 outbreak, the 2020 Summer Olympics held in Tokyo, Japan, has been rescheduled for next summer. This gives us the inspiration to look into some past Summer Olympics data and answer some questions. As a result, in this project, our primary goal is to find the correlation between the number of Summer Olympic medals each country won in 2008 and their GDP per capita, population, and sports expenditures. We also would like to answer some side questions: Who won the most medals/golden medals for his/her country? List the top ten; Which discipline has the most male/female player? Which country has the most medals/golden medals? Which year has the most female medal/golden medal winner?

For this, we collect data on the medal count, population, GDP per capita and government expenditures on sports from each country in 2008. GDP Per capita and population are obtained from the World Bank Data via the World Bank API (using URL builder string). This API requires country codes as an identifier, which is served as a part of a data set generated from R. Our medal count data set is obtained from Kaggle and loaded into MySQL database. We join it with the table of population and GDP per capita, using country codes. Unfortunately, we cannot find government expenditures on sports for all countries, so we decide to use only the available European countries' data from Eurostat and focus on European countries for this part of the analysis. We import the data in CSV form and later join it with our obtained analytical data set from previous steps. We use linear regression and scatter plots, and calculate correlations between each variable and medal count. We later choose the best fit variable (log_Population) to do a regression prediction on medal count, and find that medal count and population are positively correlated, same with GDP per capita and government expenditure, but the later two are not statistically significant at 10%.

The rest of this project is organized as follows. Section 2 shows the technical choices we made throughout the analysis. Section 3 describes the data models. Our analytics and visualization of this study are presented in section 4. Finally, our conclusion and discussion of this study are presented in Section 5.

2 Technical Choices

We choose KNIME to build our data pipeline, and it contains our complete data workflow. After eyeballing our Kaggle Summer Olympics Medals (1976-2008) data set, we realize that our "country_code" in the data set is not aligned with the ISO standard. We decide to use the World Bank API in R to get the ISO standard 2-letter country code and do some data cleaning in R before importing the data set into KNIME. This is important because we need country codes as an identifier to enable the World Bank API to get GDP per capita and population data in the next step. The screenshot (Figure 1) shows the "Load Country_Code" process in KNIME, which contains two nodes. The **R Source (Table)** node allows KNIME to read data sources from R into a KNIME table, while the **R Snippet** node executes the R script.

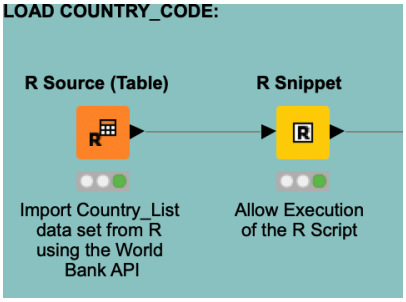


Figure 1: Workflow in KNIME (Part 1)

Next, we would like to add GDP per capita and population data into the data set by using the World Bank API. This part consists of operations to build ETL Pipeline. First, let's figure out the API we use and how we send request to it. We use The World Bank API to get the GDP per Capita and population of countries. Figure 2 gives the breakdown of the World Bank API, and this is the node **GET Request** to the API:

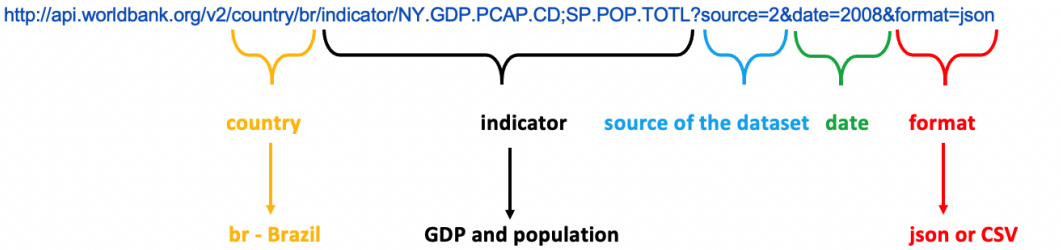


Figure 2: The World Bank API Breakdown

Here, **country/br** stands for country code, **indicator/NY.GDP.PCAP.CD;SP.POP.TOTL** stands for GDP per Capita (NY.GDP.PCAP.CD) and Population (SP.POP.TOTL) indicators, **source** stands for source of the data set, **date** is year 2008, and **format=json** means data is in JSON format.

To test the validity of request and response, we used the Postman app as well (Figure 11 in Appendix).

From Figure 3, **String Manipulation** node generates a string for this API request. This node receives *knime.in.codes* column as an input from the previous **R Snippet** node. Then, we use **join()** expression to concatenate this column and aforementioned API Breakdown to build a URL for API request:

join("http://api.worldbank.org/v2/country/",$knime.in.codes$,"/indicator/NY.GDP.PCAP.CD;SP.POP.TOTL?source=2&format=json&date=2008")

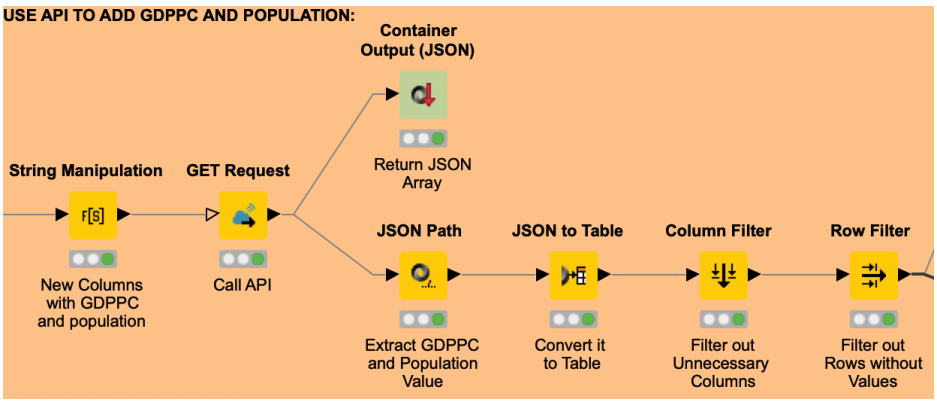


Figure 3: Workflow in KNIME (Part 2)

This node generates a **"new column"** of URL strings, which will be used for getting the GDP per Capita and population data of a specific country and year. The next **GET Request** node takes **"new column"** as an input and sends **GET Request** to the API. As a result, we get data in JSON format. **Container Output (JSON)** gives us a formatted view of the result. As an Extract part of ETL, we use **JSON Path** to extract GDP per capita and Population values out of full JSON. In the Transform part of ETL, we convert data from JSON format into table format using **JSON to Table**. Next, we use **Column Filter** to filter out unnecessary columns and **Row Filter** node to filter out rows with missing values. In the Loading of ETL, we save updated data set in a CSV file using **CSV Writer**.

Meanwhile, we load the Summer Olympics Medals (1976-2008) data set into MySQLWorkbench, and the **MySQL Connector** node creates a connection to a MySQL server via its JDBC driver. **DB Query Reader** allows us to execute the entered SQL query and returns the results as a KNIME data table. In the **DB Query Reader**, we select event_year, count of medals won and countries from the data set, limit event_year to 2008, and group all data by country. Figure 4 shows the part of workflow in KNIME. After executing this node, it gives us a table with medal count grouped by the country. We later join the table with GDP per capita and population table to have a country attributes table with the medal count.

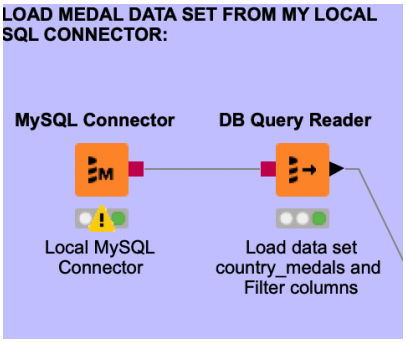


Figure 4: Workflow in KNIME (Part 3)

After joining these 2 data sets, our next step is to clean and tidy up the joined data set before any analysis work (Figure 5). We use **Row Filter** node to filter out rows with missing values. We then rename our columns using **Column Rename** node. The next node **R Snippet** is used to add log values of both GDP per capita and population to the data set. We think it will make the relationship approximate a linear pattern, as both have skewed distribution with long right tail, and there are no negative values. A part of the cleaned and tidied country attributes table is shown in Figure 10 in the Appendix. In the screenshot, we can see the table contains row ID, event_year, medal_count, country_name, country_code, gdppc, population, log_gdppc, and log_population.

Nevertheless, we cannot find government expenditures on sports for all countries, so we decide to use the available European countries' data only and focus on European countries for this part of the analysis. The data set is from Eurostat and the data is in percentage of government expenditure. We load a CSV file containing data on total government expenditure in European Countries using **File Reader**. Then we filter out unnecessary columns with **Column Filter** and only keep the following columns: year, country and total expenditure value. Since data in each year was given in separate rows, we use **Pivoting** to get each country's average total sports expenditure from 2004 to 2008. Also, we need each country in separate rows, so we use **Unpivoting** to rotate one row into many. **Cell Splitter** node is used to clean the country name column by removing unnecessary values. To keep country names and total expenditure columns, we use **Column Filter** nodes. **Column Rename** gives appropriate names to the columns and we will join them later with another table to get a complete data set. Figure 6 shows the part of the workflow.

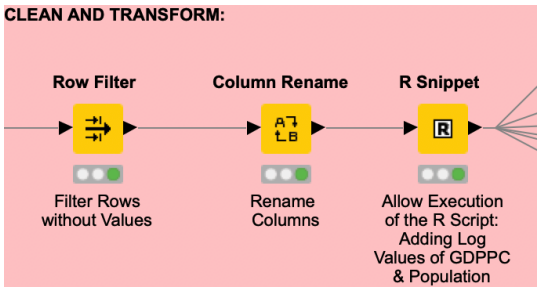


Figure 5: Workflow in KNIME (Part 4)

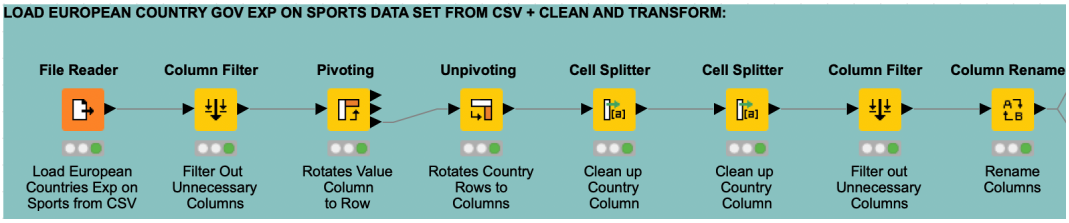


Figure 6: Workflow in KNIME (Part 5)

3 Data Models

This section demonstrates the data models we implement in this project. The following Figure 7 shows the EER diagram of our data layer. We have the original Summer Olympics Medals data set, which is joined with the country's GDP per capita and population table on country_name. It is the same with the eu_country_gov_exp_sports data set, which is joined on country_name as well. For the analytical layer (As shown in Figure 12 in Appendix), We slice the data by year (date dimension) and by country (region dimension). Finally, we use medal_count as it shows total medals each country obtained in a specific year's Summer Olympic, in our case of 2008.

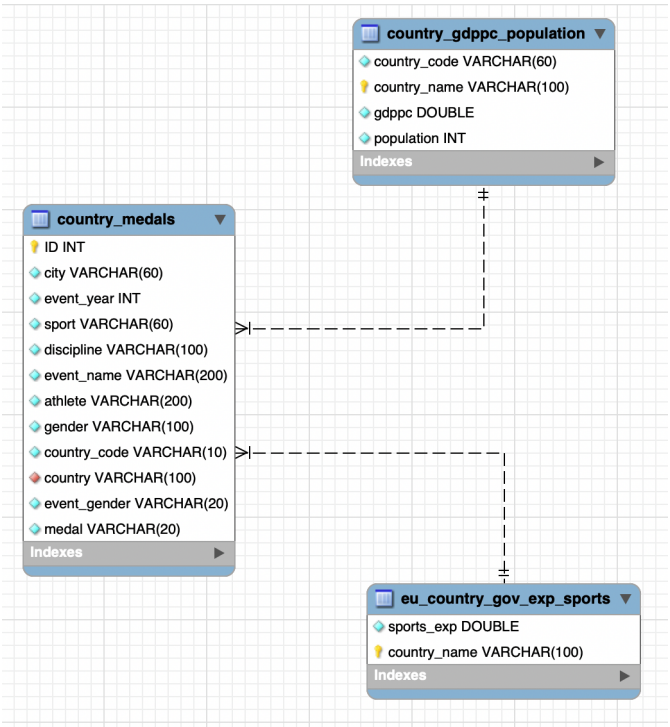


Figure 7: EER Diagram

4 Analytics and Visualization

In this section, we use the joined table to complete correlation and regression analysis between medal count and different variables (European Sports Expenditure, log form of GDP per capita and log form of the population). With the help of scatter plots, correlation values, and regression coefficients, we contrast the p value of each variable with one another, select the best-fit one, make regression prediction, and conclude which variable contributes most to the number of medals won by a country in the Olympics.

Figure 8 shows the complete Analytics part in the workflow. We use **Rank Correlation** to calculate the correlation between the number of medals won by a country and its government expenditures on sports, population and GDP per capita. To move further, we also fit a linear regression line between medal count and government expenditures on sports and between medal count and population using **Linear Regression Learner** node. Among the three, log_population seems to give the best results based on the p-values **Rank Correlation** given. The correlation value of medal count and log_population is around 0.29 with a p-value of 0.008, which is relatively small at the significance level of 0.1. Therefore, we can reject the null hypothesis and conclude that medal count and log_population are indeed positively correlated. Log_gdppc and government expenditures on sports also give a positive correlation value, but they are both statistically insignificant. Therefore, we can not reject the null, and we cannot conclude that the correlation is different from 0.

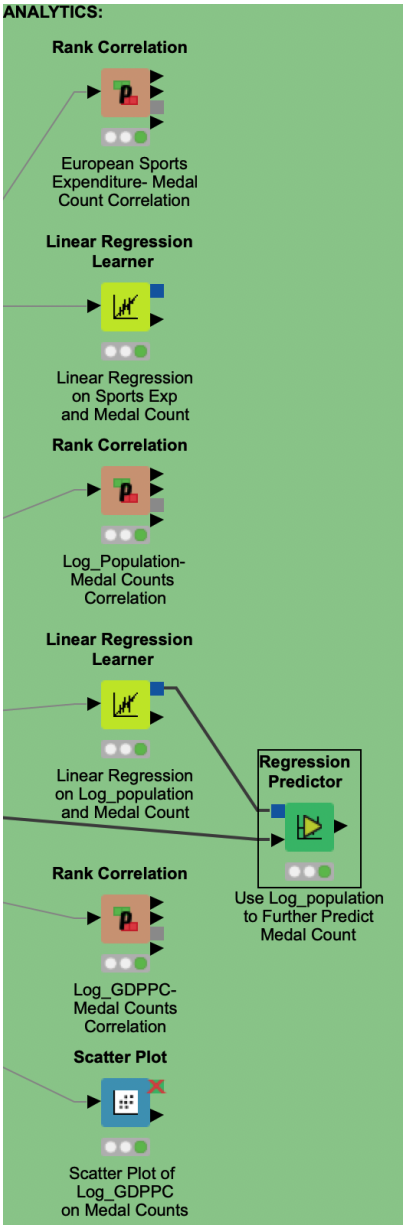


Figure 8: Workflow in KNIME (Part 5)

We also get the slope coefficient of the regression on Log_gdppc: the value of 14.386 suggests that Olympic medals earned by a country are 1.4 units higher on average for countries with ten percent higher population. We later use the node **Regression Predictor** to do a regression prediction.

To answer our side questions, we think it's best if we could visualize them in **Tableau**. We put the first two graphs to answer the first question. You can find the rest in the Appendix.

Which discipline has the most male/female player?

Based on Figure 9, Swimming events have the most female players, while Athletics(e.g running, jumping) has most male players. Also, we can find that Baseball, Boxing and Wrestling have male players only. And, softball, synchronized swimming and Rhythmic gymnastics have female players only.

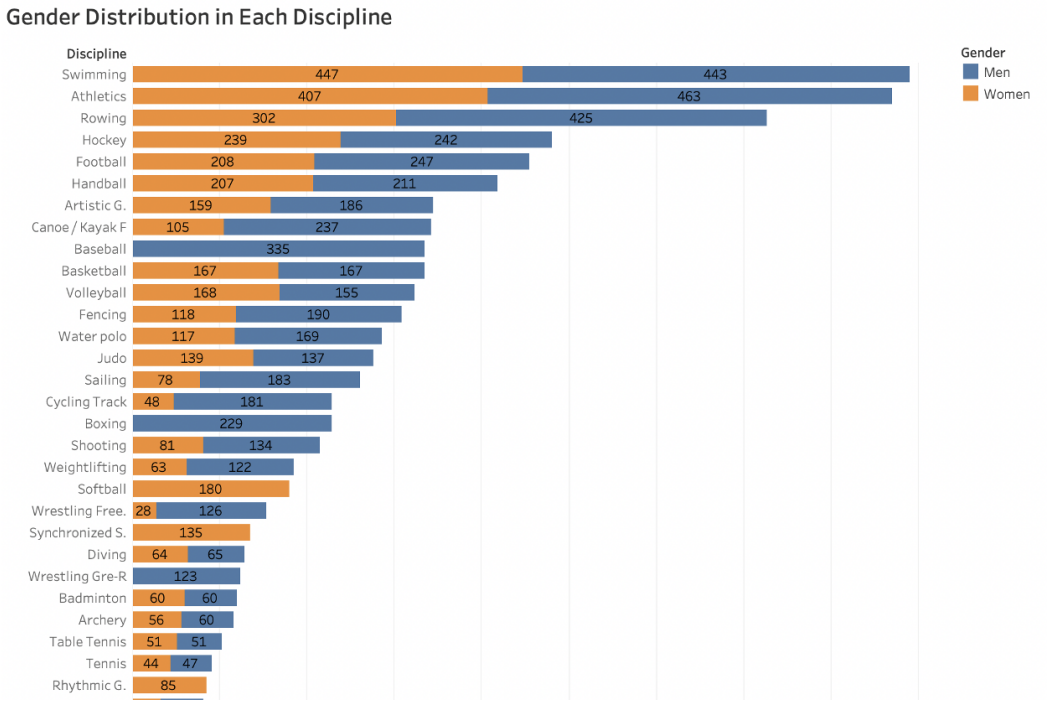


Figure 9: Bar Chart on Gender Distribution

5 Conclusion

To quickly conclude the project, we study the correlation between the number of Summer Olympic medals each country won in 2008 and their GDP per capita, population, and sports expenditures in this project. We use KNIME to document our workflow. Throughout the process, we use MySQL to build a database, use API to get data set, use R to load and clean data, and use Tableau to achieve some excellent data visualization. We find that the population is positively correlated with the medal won, and it is statistically significant at 1%. Government expenditure on sports and GDP per capita are also positively correlated with the medal won. However, the results are not statistically significant at 10%.

Below is the job descriptions on each team member:

1. **Abduvosid Malikov**: KNIME Workflow (Load Country_Code, Load European Country Gov Exp on Sports Data Set from CSV, Clean and Transform), Report (Section 3 and part of Section 2), Github Repo (Instructions on data persistence), PowerPoint Presentation

2. **Li Deborah Jia**: Tableau, KNIME Workflow (Analytics), Report (Section 4 and Appendix), Github Repo (README), PowerPoint Presentation

3. **Xinqi Wang**: KNIME Workflow (Use API to Add GDPPC and Population, Load Medal Data Set from My Local SQL Connector), Report (Section 1 and 5, part of Section 2, final editing in LaTeX), PowerPoint Presentation

A Appendix

Row ID	event_year	medal_counts	country_name	country_code	gdppc	population	log_gdppc	log_population
Row0_204	2008	315	United States	US	48,382.558	304093966	10.787	19.533
Row1_37	2008	184	China	CN	3,468.304	1324655000	8.151	21.004
Row2_11	2008	149	Australia	AU	49,601.657	21249200	10.812	16.872
Row3_165	2008	143	Russian Federation	RU	11,635.273	142742366	9.362	18.777
Row4_51	2008	101	Germany	DE	45,427.152	82110097	10.724	18.224
Row5_105	2008	78	Korea, Rep.	KR	21,350.428	49054708	9.969	17.708
Row6_69	2008	77	United Kingdom	GB	47,286.998	61806995	10.764	17.94
Row7_65	2008	76	France	FR	45,334.114	64374984	10.722	17.98
Row8_27	2008	75	Brazil	BR	8,831.023	192030362	9.086	19.073
Row9_60	2008	71	Spain	ES	35,366.26	45954106	10.474	17.643
Row10_144	2008	62	Netherlands	NL	57,644.48	16445593	10.962	16.616
Row11_7	2008	51	Argentina	AR	9,020.873	40080160	9.107	17.506
Row12_98	2008	51	Japan	JP	39,339.298	128063000	10.58	18.668
Row13_46	2008	47	Cuba	CU	5,411.271	11236971	8.596	16.235
Row14_95	2008	42	Italy	IT	40,778.343	58826731	10.616	17.89
Row15_33	2008	34	Canada	CA	46,594.451	33247118	10.749	17.319
Row16_202	2008	31	Ukraine	UA	3,887.242	46258189	8.265	17.65
Row18_86	2008	27	Hungary	HU	15,753.473	10038188	9.665	16.122
Row19_142	2008	24	Nigeria	NG	2,242.872	150269623	7.716	18.828
Row20_145	2008	22	Norway	NO	96,944.096	4768212	11.482	15.377
Row21_164	2008	22	Romania	RO	10,435.044	20537875	9.253	16.838
Row22_156	2008	20	Poland	PL	14,001.382	38125759	9.547	17.456
Row23_54	2008	18	Denmark	DK	64,322.064	5493621	11.072	15.519
Row24_96	2008	17	Jamaica	JM	4,917.718	2781876	8.501	14.839
Row25_176	2008	15	Serbia	RS	7,101.04	7350222	8.868	15.81
Row26_93	2008	14	Iceland	IS	56,409.773	317414	10.94	12.668
Row27_100	2008	14	Kenya	KE	902.07	39791981	6.805	17.499
Row28_148	2008	14	New Zealand	NZ	31,290.254	4259800	10.351	15.265
Row29_99	2008	13	Kazakhstan	KZ	8,513.565	15674000	9.049	16.568
Row30_34	2008	11	Switzerland	CH	72,487.846	7647675	11.191	15.85
Row31_180	2008	10	Slovak Republic	SK	18,677.293	5379233	9.835	15.498
Row32_197	2008	8	Turkey	TR	10,854.172	70418604	9.292	18.07
Row33_62	2008	7	Ethiopia	ET	326.437	82916235	5.788	18.233
Row34_87	2008	7	Indonesia	ID	2,166.854	235469762	7.681	19.277
Row35_77	2008	7	Greece	GR	31,997.282	11077841	10.373	16.22
Row36_182	2008	7	Sweden	SE	56,152.552	9219637	10.936	16.037
Row37_50	2008	7	Czech Republic	CZ	22,698.854	10384603	10.03	16.156

Figure 10: Part of Cleaned Country Attributes Table

Continue with Visualization: Some Fun Facts

2. Who won the most medals/golden medals for his/her country? (Figure 13)

Phelps Michael leads the first in both medal and gold medal rankings. The second is Thompson Jenny, who won 8 gold and 12 medals in total. Also, most leading sports players are from the U.S.A.

3. Which country has the most medals/golden medals? (Figure 14)

The U.S.A leads first in both rankings. In medal ranking, the second and third are Australia and Russia, while in gold medal ranking, they are China and Russia.

4. Which year has the most female medal/golden medal winner? (Figure 15)

There is a rising trend for both male and female winners. Within the timeframe of 1992 to 2008, the last year(2008) has most female medal/golden medal winners.

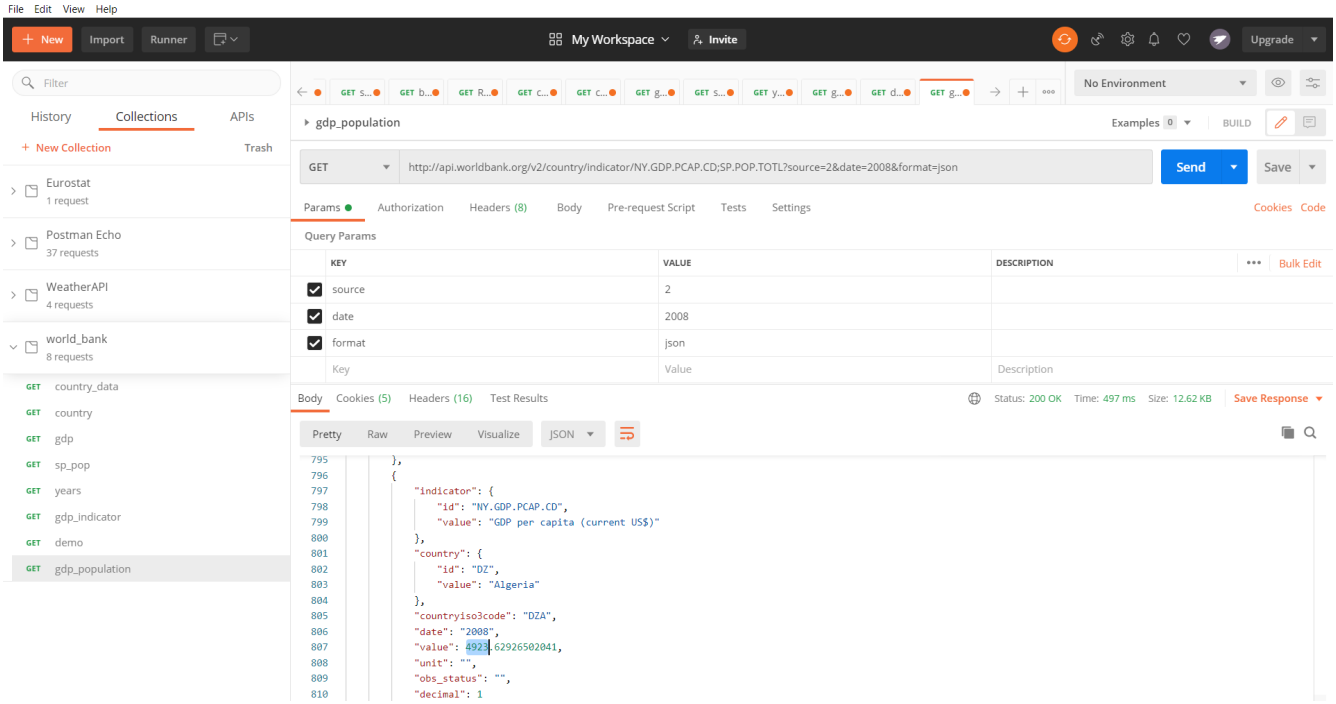
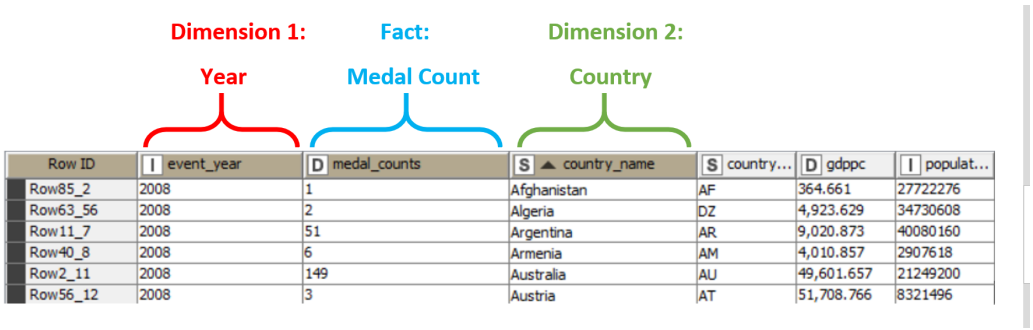
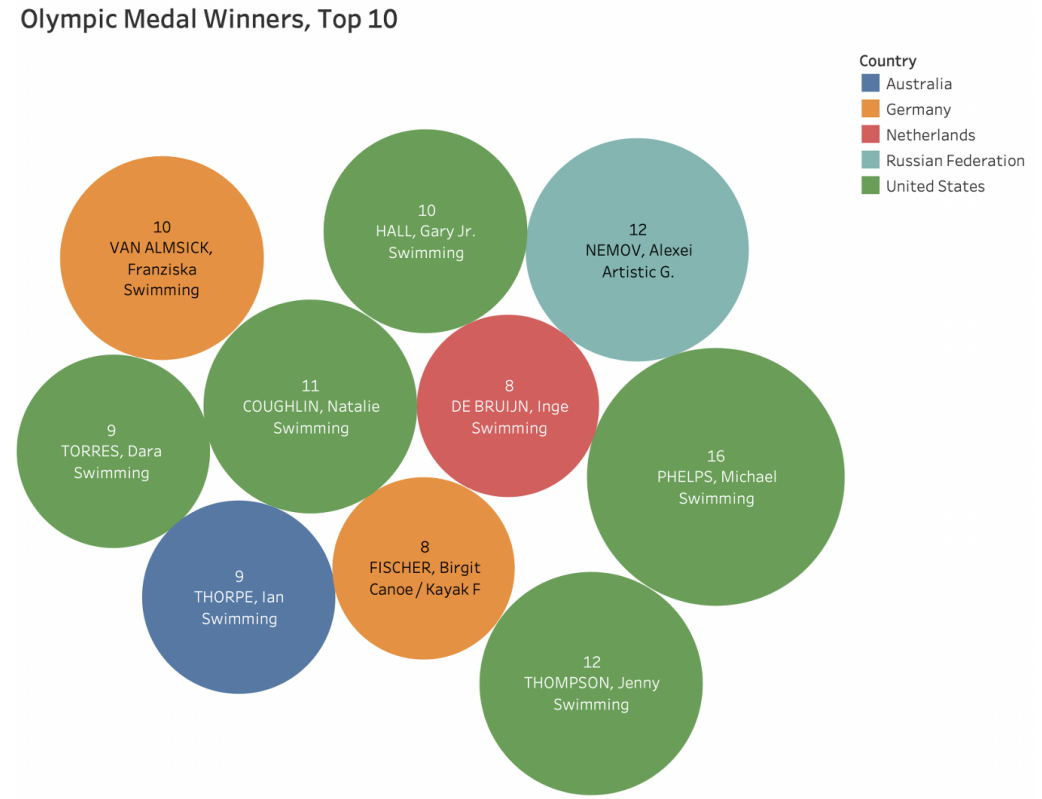


Figure 11: Screenshot of Postman

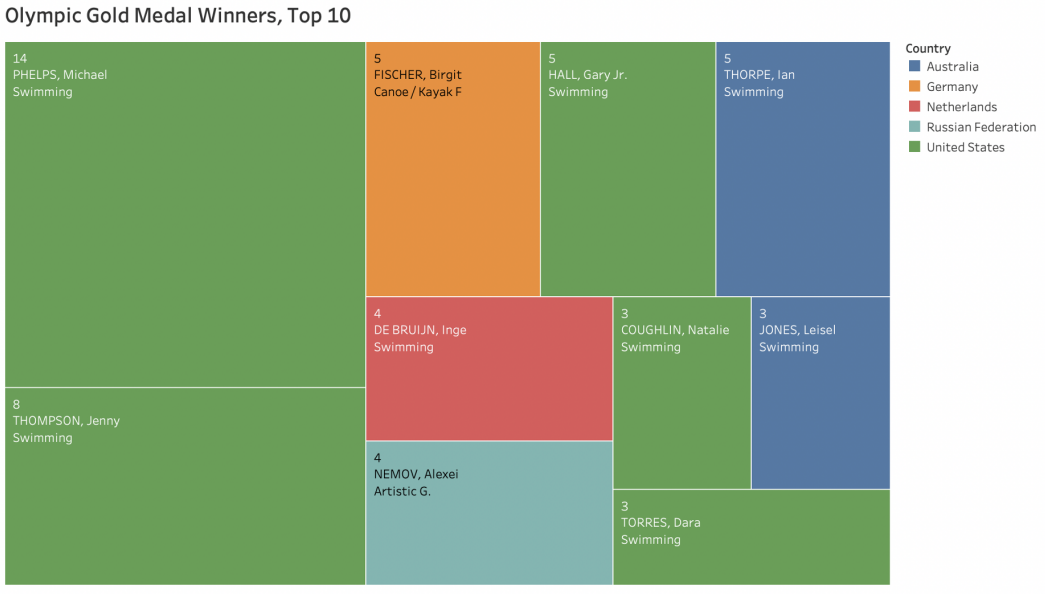


Row ID	event_year	medal_counts	country_name	country...	gdppc	populat...
Row85_2	2008	1	Afghanistan	AF	364.661	27722276
Row63_56	2008	2	Algeria	DZ	4,923.629	34730608
Row11_7	2008	51	Argentina	AR	9,020.873	40080160
Row40_8	2008	6	Armenia	AM	4,010.857	2907618
Row2_11	2008	149	Australia	AU	49,601.657	21249200
Row56_12	2008	3	Austria	AT	51,708.766	8321496

Figure 12: Analysis Diagram

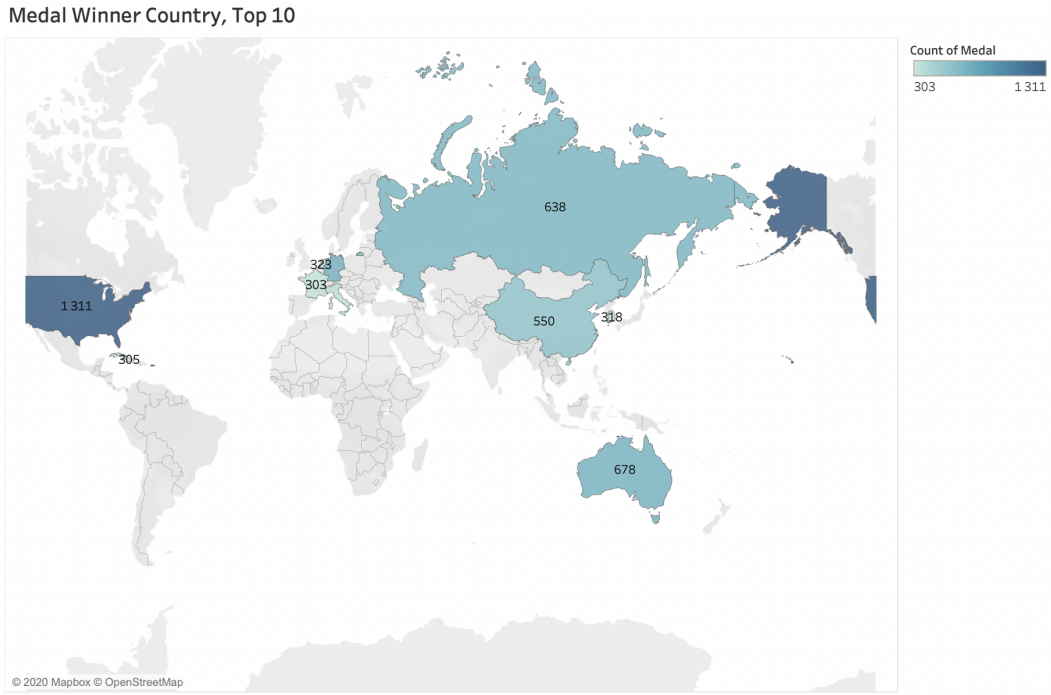


(a) Bubble Chart

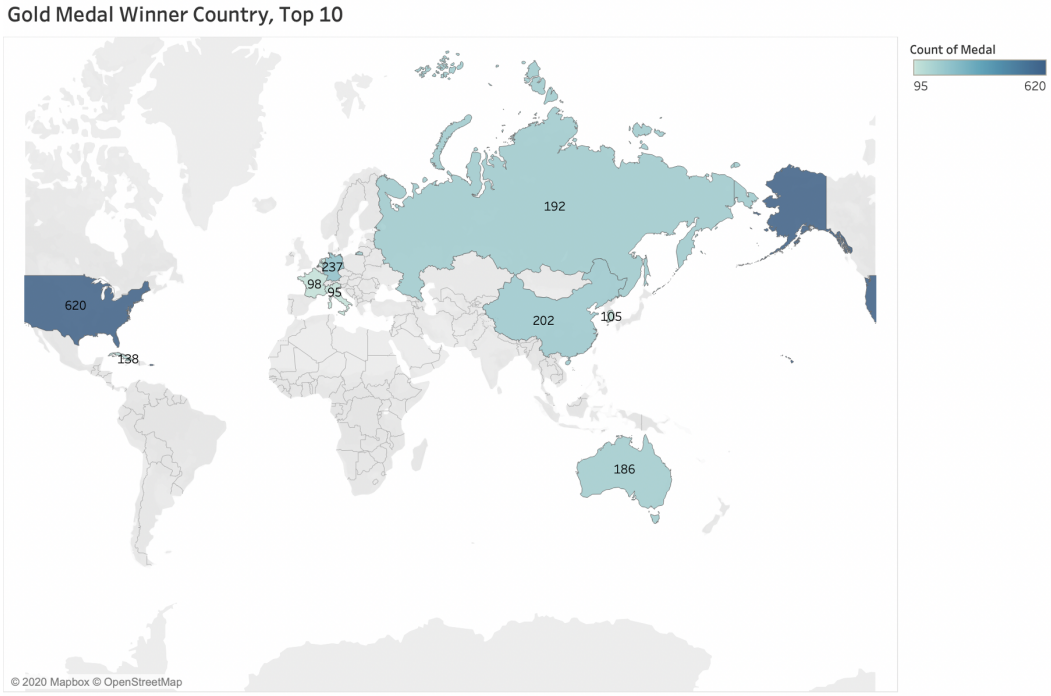


(b) Tree Map

Figure 13: Bubble Chart and Tree Map

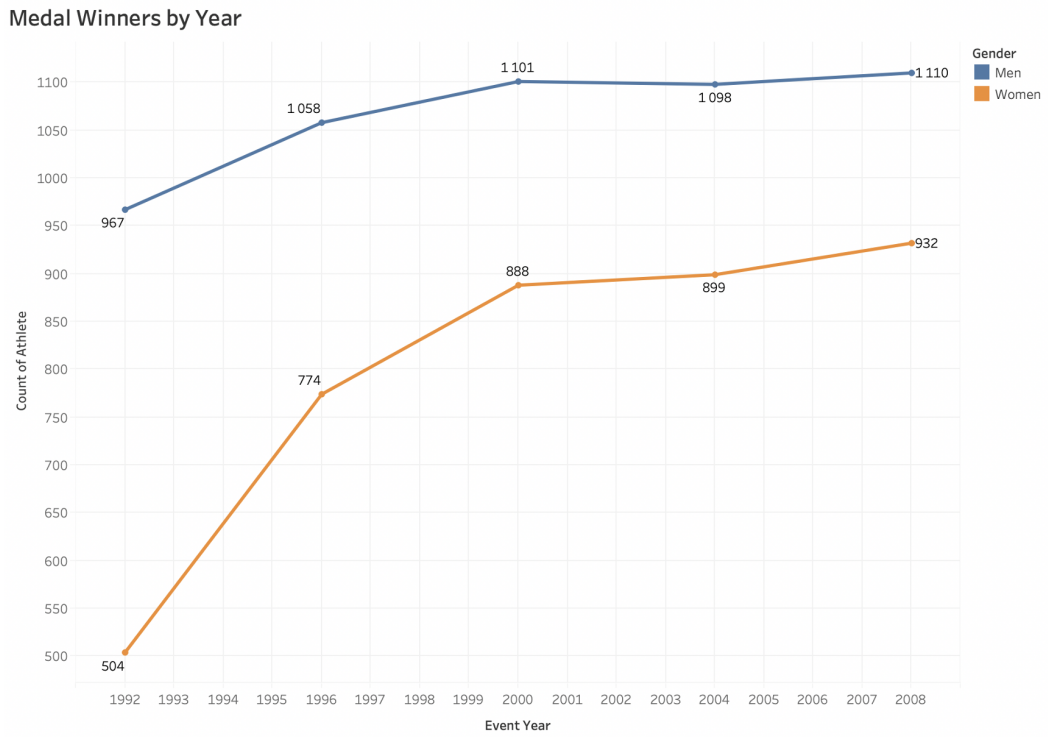


(a) World Map with Medal Winner Country, Top 10

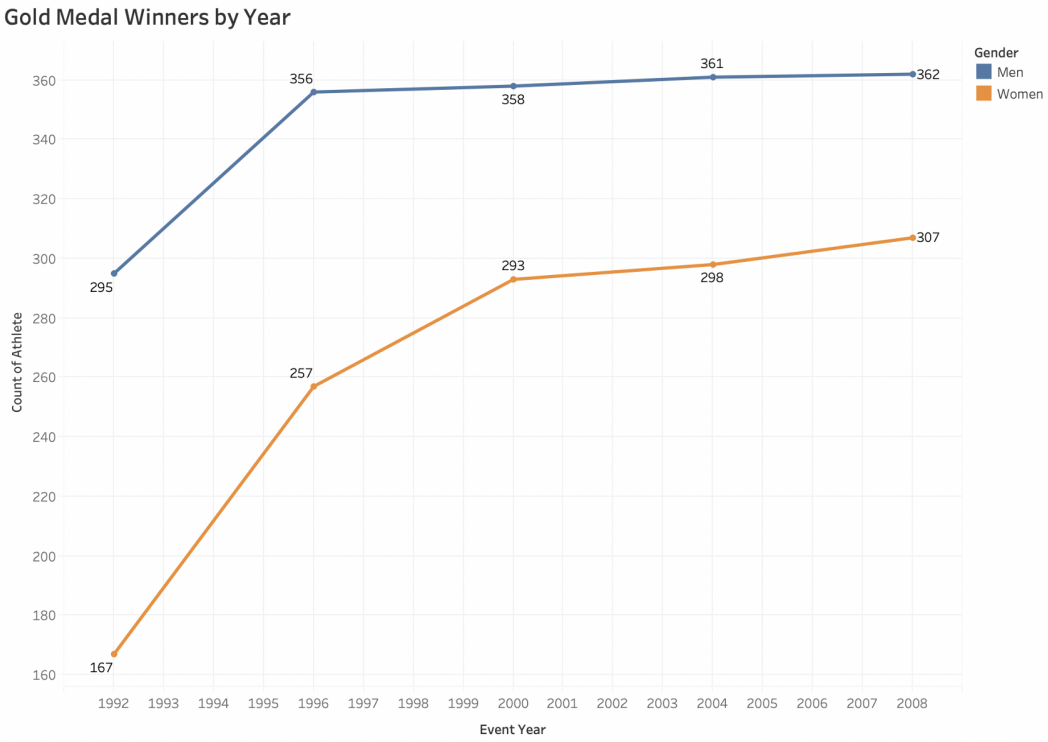


(b) World Map with Gold Medal Winner Country, Top 10

Figure 14: World Maps



(a) Medal Winner by Year Line Chart



(b) Gold Medal Winner by Year Line Chart

Figure 15: Line Charts