Information Visualization

CHECKPOINT II: Data cleaning and processing

G01-A

Changes: In the table in the second page, we added the file names of each picture shown.

1. Initial Dataset

We have a table (All Winners) containing all the editions of the games from 1896 until 2008, and for each there is the city and year they were at, and a set of podium finishes, which include the sport, discipline and event, and the athlete's name and country code (NOC, same as IOC and 3-letter ISO) and the medal they won. We have another table (Total) with the each country's name, NOC, total medals and how many of gold, silver and bronze.

We have a table (Codes) with each country name, IOC (same as NOC) code and their ISO code.

We have another table (Population) with each country's name, country code (3-letter ISO), indicator name and code, and a field of their population for each year from 1960 until 2014.

We have a table (Coordinates) to know the coordinates of each country, which has its 2-letter ISO code, latitude and longitude, and its name.

2. Selected/Derived Data

All Winners – Edition year, Sport, Medal won and the NOC of the country of the medallist.
Total – NOC, Total of medals.
Codes – Country, IOC, ISO code.
Population – ISO Code, Years from 1960 until 2008.
Coordinates – ISO Code, Latitude, Longitude.
Derived measure (*(medals won)/(population)* coefficient) – We want to compare the medals per capita over the years, so we will count the number of medals each country won in each year and divide it by its population in that year.

3. Data abstraction

All Winners – A tree containing all the podium finishes of the countries since 1896 until 2008. It's organized first by **edition** which is the year of the games, then by **sport** which can be "swimming", "athletics", etc. and finally, **the number of medals won** which can be represented as "**Bronze**", "**Silver**" or "**Gold**" , and the **NOC** of the medallist.
Total – A simple table with the **NOC**, which is a 3-letter code representing a country; and an integer which is **the total count of medals** from all the editions of the games.
Codes – A simple table with sets of three strings: the **country name**, the **IOC** (country code, equal to NOC) and a **2-letter ISO** country code.
Population – A table with a **3-letter ISO** country code matching the **IOC**, and a set of columns each pertaining **to every fourth year between 1960 and 2008**, containing the **population** of the country in that year.
Coordinates – A table with a **2-letter ISO** country code for each **country** (*nominal*), and a **latitude** and **longitude** for that country.

Attribute	Type	Semantics
Edition	Quantitative (Continuous)	Independent, Discrete, Temporal, Non Spatial
Sport	Nominal	Dependent, Discrete, Temporal, Non Spatial
NOC	Nominal	Dependent, Discrete, Temporal, Spatial
Medal	Ordinal	Dependent, Discrete, Temporal, Non Spatial

Total	Quantitative (Ratio)	Independent, Discrete, Non Temporal, Non Spatial
Gold	Quantitative (Ratio)	Independent, Discrete, Non Temporal, Non Spatial
Attribute	**Type**	**Semantics**
Silver	Quantitative (Ratio)	Independent, Discrete, Non Temporal, Non Spatial
Bronze	Quantitative (Ratio)	Independent, Discrete, Non Temporal, Non Spatial
Country	Nominal	Independent, Discrete, Non Temporal, Spatial
ISO	Nominal	Independent, Discrete, Non Temporal, Spatial
Years	Quantitative (Ratio)	Dependent, DiscreteTemporal, Non Spatial
Latitude	Quantitative (Continuous)	Dependent, Continuous, Non Tempora, Spatial
Longitude	Quantitative (Continuous)	Dependent, Continuous, Non Tempora, Spatial

4. Dataset processing

To create the Codes table, we had to make sure the IOC codes matched the NOC on All Winners and Totals, and that the 2 and 3 letter ISO codes matched the same on other tables. Some values didn't exist, because they were for older countries or united teams, so we checked the most representative countries related to those and made the association.
We used Pentaho's Group By to sort the All Winners table according to various attributes and sum values to get the totals for the amounts of medals over time.
To get the *(medals won)/(population)* coefficient, we counted the amount of medals for each country (Pentaho's *Group By* on All Winners) and divided it by the population of that country in that year (using Pentaho's *Merge Join* between the two tables, then *Calculation*).

5. Mapping (Data sample / Questions)

1 – What countries had the most gold medalists in the first games, in 1896?
Year, Code, Medal, Amount

1896	GER	Gold	26
1896	GER	Silver	5
1896	GRE	Bronze	22
1896	GRE	Gold	10
1896	GRE	Silver	20
1896	HUN	Bronze	3

vitorias por medalha ano.csv

2 – What country has the most medallists in Judo?
Sport, Code, Amount

Judo	CUB	32
Judo	FRA	37
Judo	KOR	37
Judo	JPN	65
Lacrosse	USA	13

vitorias totais por desporto.csv

3 – What are the standings of the USSR in 1964?
Year, Code, Medal, Amount

1964	TUR	Bronze	1
1964	TUR	Gold	2
1964	TUR	Silver	3
1964	URS	Bronze	50
1964	URS	Gold	61
1964	URS	Silver	63
1964	URU	Bronze	1
1964	USA	Bronze	36

vitorias por medalha ano.csv

4 – See the countries with the most medallists per capita in 2008.
Victories, Country, Code, Year, Population, Coefficient*1 000 000

1	Channel Is	CHI	2008	157,587.00	6.345701
17	Jamaica	JAM	2008	2,671,934.00	6.362433
149	Australia	AUS	2008	21,249,200.00	7.012029
5	Bahamas,	BAH	2008	348,587.00	14.34362
14	Iceland	ISL	2008	317,414.00	44.10644

test coef 0.csv

5 – How do the USSR and Russia's cumulative scores compare?

Country	NOC CODE	Total
USA	USA	2297
Soviet Uni	URS	1010
Germany (	GER	851
Great Brit	GBR	714
France	FRA	638
Italy	ITA	521
Sweden	SWE	475
Hungary	HUN	458
Australia	AUS	432
East Germ	FRG	409
China	CHN	385
Japan	JPN	360
Russia	RUS	317

fixed totals alltime.csv