

Information Visualization

Project Proposal and Dataset



70493 – Tiago Nascimento

G01-A

76102 – Miguel Cruz

76394 – Daniel Trindade

01

INITIAL DATASET

Initial Dataset

Description

All Winners – contains the podium finishers of all time, for each edition and for each sport

Data sample

City	Edition	Sport	Discipline	Athlete	NOC	Gender	Event	Event_gender	Medal
Athens	1896	Aquatics	Swimming	HAJOS, Alfred	HUN	Men	100m freestyle	M	Gold

Initial Dataset

Description

Population – contains the country name, their ISO codes (with 3 letters) and their population between 1960 and 2014.

Data sample

Country Name	Country Code	Indicator Name	Indicator Code	1960
Aruba	ABW	Population, total	SP.POP.TOTL	54208

...

Initial Dataset

Description

Total – contains the country name, their NOC code and the medals won since the beginning of the Olympic Games

Data sample

Country	NOC CODE	Total	Golds	Silvers	Bronzes
USA	USA	2297	930	728	639

Initial Dataset

Description

Codes – contains the 2-letter ISO and IOC codes from all countries.

Data sample

Country	Int Olympic Committee code	ISO code	Country
Afghanistan	AFG	AF	Afghanistan

Initial Dataset

Description

Coordinates – contains the 2-letter ISO code of a country, their coordinates and the names of the country.

Data sample

country	Latitude	Longitude	name
AD	33	65	Andorra

02

**SELECTED / DERIVED
DATA**

Derived data

Derived data description

Compare the medals per capita over the years:

- 1) Count the medals each country won in each year
- 2) Know the population of a country in that year

Derived data

Why?

Why?

Answer:

Is it possible that, the more population a country has, the more probability is to have more winners in the olympic games?

Selected data

Data description

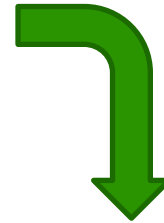
- 1) From the “All Winners” dataset, we wanted:
 - The OG edition;
 - The Sport;
 - The winning medal
 - The NOC code

Selected data

Data description

1) From the “All Winners” dataset, we wanted:

- The OG edition;
- The Sport;
- The winning medal
- The NOC code



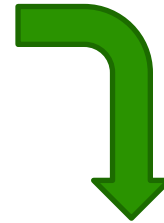
Edition	Sport	NOC	Medal
1896	Aquatics	HUN	Gold
1896	Aquatics	AUT	Silver
1896	Aquatics	GRE	Bronze

Selected data

Data description

1) From the “All Winners” dataset, we wanted:

- The OG edition;
- The ~~event~~ sport;
- The winning medal
- The NOC code



Edition	NOC	Medal
1896	HUN	Gold
1896	AUT	Silver
1896	GRE	Bronze

Selected data

Data description

- 2) From the “Population” dataset, we wanted:
- The ISO Code,
 - Years (that contained the population)

Selected data

Data description

2) From the “Population” dataset, we wanted:

- The ISO Code,
- Years (that contained the population)

Country Name	Country Code	1960	1961
Aruba	ABW	54208	55435
Andorra	AND	13414	14376
Afghanistan	AFG	8994793	9164945
Angola	AGO	5270844	5367287
Albania	ALB	1608800	1659800
Arab World	ARB	92495902	95041593
United Arab Emirates	ARE	92612	100985

Our Result...

For the year 1960:

$$X = \frac{\text{Total Medals of a Country in 1960}}{\text{Population of the Country in 1960}}$$

03

DATA ABSTRACTION

Data abstraction

Description:


All Winners – A table containing all the podium finishes of the countries since 1896 until 2008.

Dataset type:

Organized as a tree, first by the year, then the sport and then the NOC country code and the medal.

Data abstraction (attribute types)

Edition	Sport	NOC	Medal
1896	Aquatics	HUN	Gold
1896	Aquatics	AUT	Silver
1896	Aquatics	GRE	Bronze
1896	Aquatics	GRE	Gold
1896	Aquatics	GRE	Silver
1896	Aquatics	GRE	Bronze
1896	Aquatics	HUN	Gold
1896	Aquatics	GRE	Silver



Quantitative
(Continuous)



Nominal

Data abstraction (semantics)

Edition	Sport	NOC	Medal
1896	Aquatics	HUN	Gold
1896	Aquatics	AUT	Silver
1896	Aquatics	GRE	Bronze
1896	Aquatics	GRE	Gold
1896	Aquatics	GRE	Silver
1896	Aquatics	GRE	Bronze
1896	Aquatics	HUN	Gold
1896	Aquatics	GRE	Silver

Independent
Discrete
Temporal
Non Spatial

Dependent
Discrete
Temporal
Non Spatial

Dependent
Discrete
Temporal
Spatial

Dependent
Discrete
Temporal
Non Spatial

Data abstraction

Description:

Total – A table containing the total number of medals won and for each type of medal.

Dataset type:

Organized as a tree, first by the country, then its NOC code, then by total Medals won, and then Medal specific (Gold,Silver,Bronze)

Data abstraction (attribute types)

Country	NOC CODE	Total	Gold	Silver	Bronze
USA	USA	2297	930	728	639
Soviet Union	URS	1010	395	319	296
Germany (incl)	GER	851	247	284	320
Great Britain	GBR	714	207	255	252
France	FRA	638	192	212	234
Italy	ITA	521	190	157	174
Sweden	SWE	475	142	160	173
Hungary	HUN	458	159	140	159
Australia	AUS	432	131	137	164

Nominal

Quantitative
(Ratio)

Data abstraction (semantics)

Country	NOC CODE	Total	Gold	Silver	Bronze
USA	USA	2297	930	728	639
Soviet Union	URS	1010	395	319	296
Germany (incl)	GER	851	247	284	320
Great Britain	GBR	714	207	255	252
France	FRA	638	192	212	234
Italy	ITA	521	190	157	174
Sweden	SWE	475	142	160	173
Hungary	HUN	458	159	140	159
Australia	AUS	432	131	137	164

Independent
Discrete
Non Temporal
Spatial

Independent
Discrete
Non Temporal
Non Spatial

Data abstraction

Description:


Codes – A table containing all countries and for each its IOC and its 2letter ISO code

Dataset type:

Sets of 3 strings Country, IOC , 2 letter ISO code

Data abstraction (attribute types)


Country	Int Olympic Comm	ISO code
Afghanistan	AFG	AF
Albania	ALB	AL
Algeria	ALG	DZ
American Samoa*	ASA	AS
Andorra	AND	AD
Angola	ANG	AO
Antigua and Barbuda	ANT	AG
Argentina	ARG	AR
Armenia	ARM	AM
Aruba*	ARU	AW
Australia	AUS	AU



Nominal

Data abstraction (semantics)

Country	Int Olympic Commi	ISO code
Afghanistan	AFG	AF
Albania	ALB	AL
Algeria	ALG	DZ
American Samoa*	ASA	AS
Andorra	AND	AD
Angola	ANG	AO
Antigua and Barbuda	ANT	AG
Argentina	ARG	AR
Armenia	ARM	AM
Aruba*	ARU	AW
Australia	AUS	AU



Independent
Discrete
Non Temporal
Spatial

Data abstraction

Description:

Population – A table that for each country has its population since 1960.

Dataset type: 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.

Data abstraction (attribute types)

Country Name	Country Code	1960	1961	1962	1963	1964	1965
Aruba	ABW	54208	55435	56226	56697	57029	57360
Andorra	AND	13414	14376	15376	16410	17470	18551
Afghanistan	AFG	8994793	9164945	9343772	9531555	9728645	9935358
Angola	AGO	5270844	5367287	5465905	5565808	5665701	5765025
Albania	ALB	1608800	1659800	1711319	1762621	1814135	1864791
Arab World	ARB	92495902	95041593	97691498	100438281	103273929	106192292
United Arab Emirates	ARE	92612	100985	112240	125216	138220	150318
United Arab Emirates	UAE	92612	100985	112240	125216	138220	150318
Argentina	ARG	20619075	20953079	21287682	21621845	21953926	22283389
Armenia	ARM	1867396	1934239	2002170	2070427	2138133	2204650
American Samoa	ASM	20012	20478	21118	21883	22701	23518

Nominal

Quantitative
(Ratio)

Data abstraction (semantics)

Country Name	Country Code	1960	1961	1962	1963	1964	1965
Aruba	ABW	54208	55435	56226	56697	57029	57360
Andorra	AND	13414	14376	15376	16410	17470	18551
Afghanistan	AFG	8994793	9164945	9343772	9531555	9728645	9935358
Angola	AGO	5270844	5367287	5465905	5565808	5665701	5765025
Albania	ALB	1608800	1659800	1711319	1762621	1814135	1864791
Arab World	ARB	92495902	95041593	97691498	100438281	103273929	106192292
United Arab Emirates	ARE	92612	100985	112240	125216	138220	150318
United Arab Emirates	UAE	92612	100985	112240	125216	138220	150318
Argentina	ARG	20619075	20953079	21287682	21621845	21953926	22283389
Armenia	ARM	1867396	1934239	2002170	2070427	2138133	2204650
American Samoa	ASM	20012	20478	21118	21883	22701	23518

Independent
Discrete
Non Temporal
Spatial

Dependent
Discrete
Temporal
Non Spatial

Data abstraction

Description:

Coordinates – A table that for each country has its longitude and latitude coordinates.

Dataset type: A table with a **2-letter ISO** country code for each country and a latitude and longitude for that country.

Data abstraction (attribute types)

country	latitude	longitude	name
AD	42.546245	1.601554	Andorra
AE	23.424076	53.847818	United Arab Emirates
AF	33.93911	67.709953	Afghanistan
AG	17.060816	-61.796428	Antigua and Barbuda
AI	18.220554	-63.068615	Anguilla
AL	41.153332	20.168331	Albania
AM	40.069099	45.038189	Armenia
AN	12.226079	-69.060087	Netherlands Antilles
AO	-11.202692	17.873887	Angola

Nominal

Quantitative
(Continuous)

Nominal

Data abstraction (semantics)

country	latitude	longitude	name
AD	42.546245	1.601554	Andorra
AE	23.424076	53.847818	United Arab Emirates
AF	33.93911	67.709953	Afghanistan
AG	17.060816	-61.796428	Antigua and Barbuda
AI	18.220554	-63.068615	Anguilla
AL	41.153332	20.168331	Albania
AM	40.069099	45.038189	Armenia
AN	12.226079	-69.060087	Netherlands Antilles
AO	-11.202692	17.873887	Angola

Independent
Discrete
Non Temporal
Spatial

Dependent
Continuous
Non Temporal
Spatial

Independent
Discrete
Non Temporal
Spatial

04

DATASET PROCESSING

Dataset processing

Dataset cleaning description

- Did all IOC values existed?
- Were the IOC codes the same as NOC codes?
- What to do with ISO codes with 3 letters?
- (Parte do Miguel)

Dataset processing

Problem found

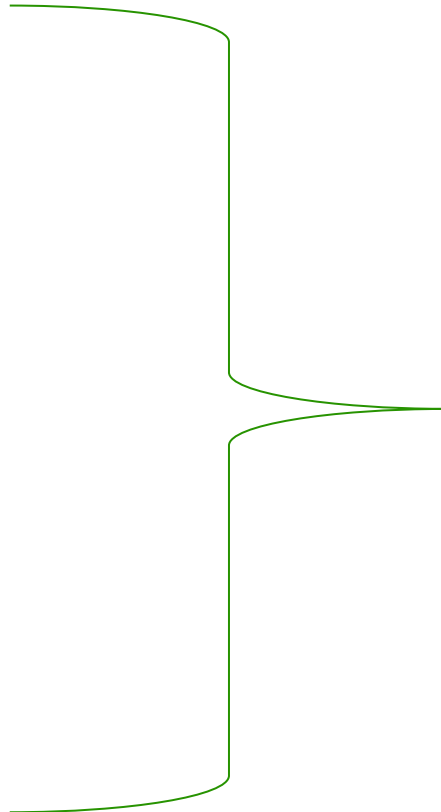
IOC Values didn't exist



Dataset processing

Problem found

IOC Values don't exist anymore



Dataset processing

Solution for Problem 1:

Gave the medals they won to the actual
Germany



Germany	GER	DE	Germany
Germany	EUA	DE	Germany
Germany	FRG	DE	Germany
Germany	GDR	DE	Germany

Dataset processing

Problem found

IOC = NOC ?



Dataset processing

Problem found

IOC = NOC ?



IOC Code -> ROM

NOC Code -> ROU

Dataset processing

Solution for problem 2

Give to the IOC attribute, the values of NOC and IOC values

Romania	ROM	RO	Romania
Romania	ROU	RO	Romania



They were always the same!

Dataset processing

Problem found

3-letter ISO Codes Conflite



ISO Code -> PRT

Dataset processing

Solution for problem 3

Grant to Portugal, the IOC code, to prevent confusions.

Portugal	PRT	8857716	8929316	8993985	9030355
Portugal	POR	8857716	8929316	8993985	9030355

Dataset processing

How to get the winners/population coefficient?

We needed the amount of medals for each country, each year...

Edition	NOC	Medal
1896	HUN	Gold
1896	AUT	Silver
1896	GRE	Bronze



Group By

Edition	NOC	Vitorias por ano
1896	AUS	2
1896	AUT	5
1896	DEN	6
1896	FRA	11
1896	GBR	7
1896	GER	33
1896	GRE	52

Dataset processing

How to get the winners/population coefficient?

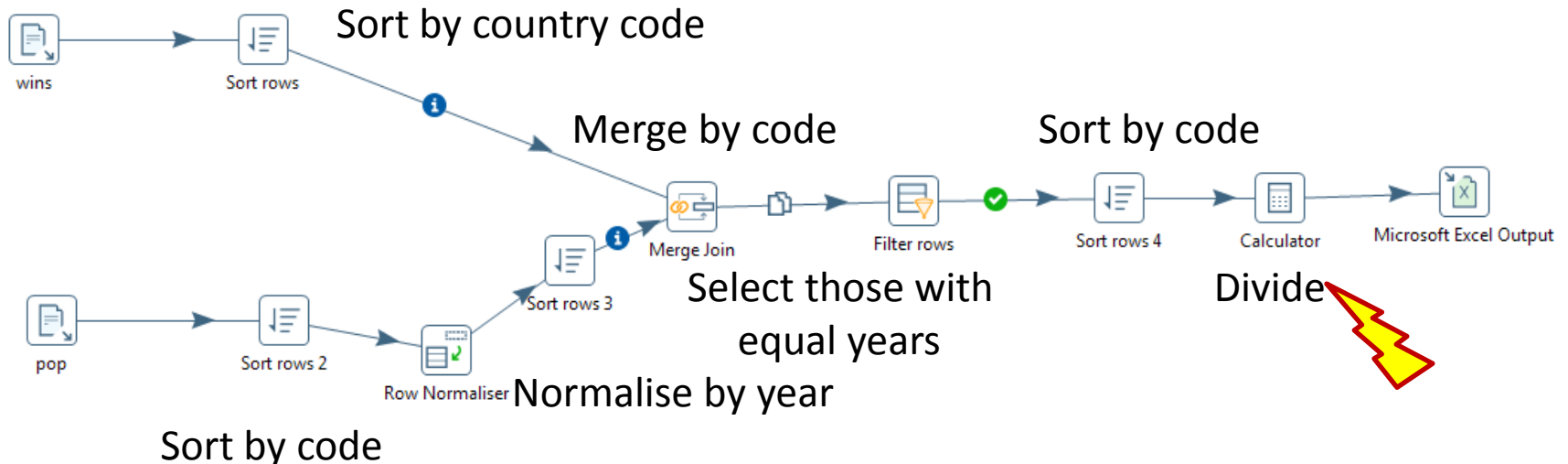
We also had the population

Country Name	Country Code	1960	1961
Aruba	ABW	54208	55435
Andorra	AND	13414	14376
Afghanistan	AFG	8994793	9164945
Angola	AGO	5270844	5367287
Albania	ALB	1608800	1659800
Arab World	ARB	92495902	95041593
United Arab Emirates	ARE	92612	100985

Dataset processing

How to get the winners/population coefficient?

We let Pentaho work on that...



Dataset processing

How to get the winners/population coefficient?

We used Excel to calculate with the needed precision and we made the results more readable:

Vitorias po	Country Na	Code	Year	Amount	Coeficiente
1.00	Afghanista	AFG	2008	26,528,741.00	0.037695
1.00	Netherland	AHO	1988	14,760,094.00	0.0677502
2.00	Algeria	ALG	1984	21,893,857.00	0.0913498
2.00	Algeria	ALG	1992	27,180,921.00	0.073581

Medalists/Population x 1 000 000

05

MAPPING

Mapping

1 – What countries had the most gold medallists in the first games, in 1896?

Edition	NOC	Medal	Vitorias por medalha
1896	GER	Gold	26
1896	GRE	Bronze	22
1896	GRE	Silver	20
1896	USA	Gold	11
1896	GRE	Gold	10
1896	USA	Silver	7

2 – What country has the most medallists in Judo?

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

3 – What are the standings of the USSR in 1964?

1964	TUR	Silver	1
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

Mapping

4 – See the countries with the most medallists per capita in 2008.

14,00	Iceland	ISL	2008	317 414,00	44,106435
5,00	Bahamas,	BAH	2008	348 587,00	14,343622
149,00	Australia	AUS	2008	21 249 200,00	7,0120287
17,00	Jamaica	JAM	2008	2 671 934,00	6,3624326
1,00	Channel Is	CHI	2008	157 587,00	6,3457011
22,00	Norway	NOR	2008	4 768 212,00	4,6138888
47,00	Cuba	CUB	2008	11 290 239,00	4,162888

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

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