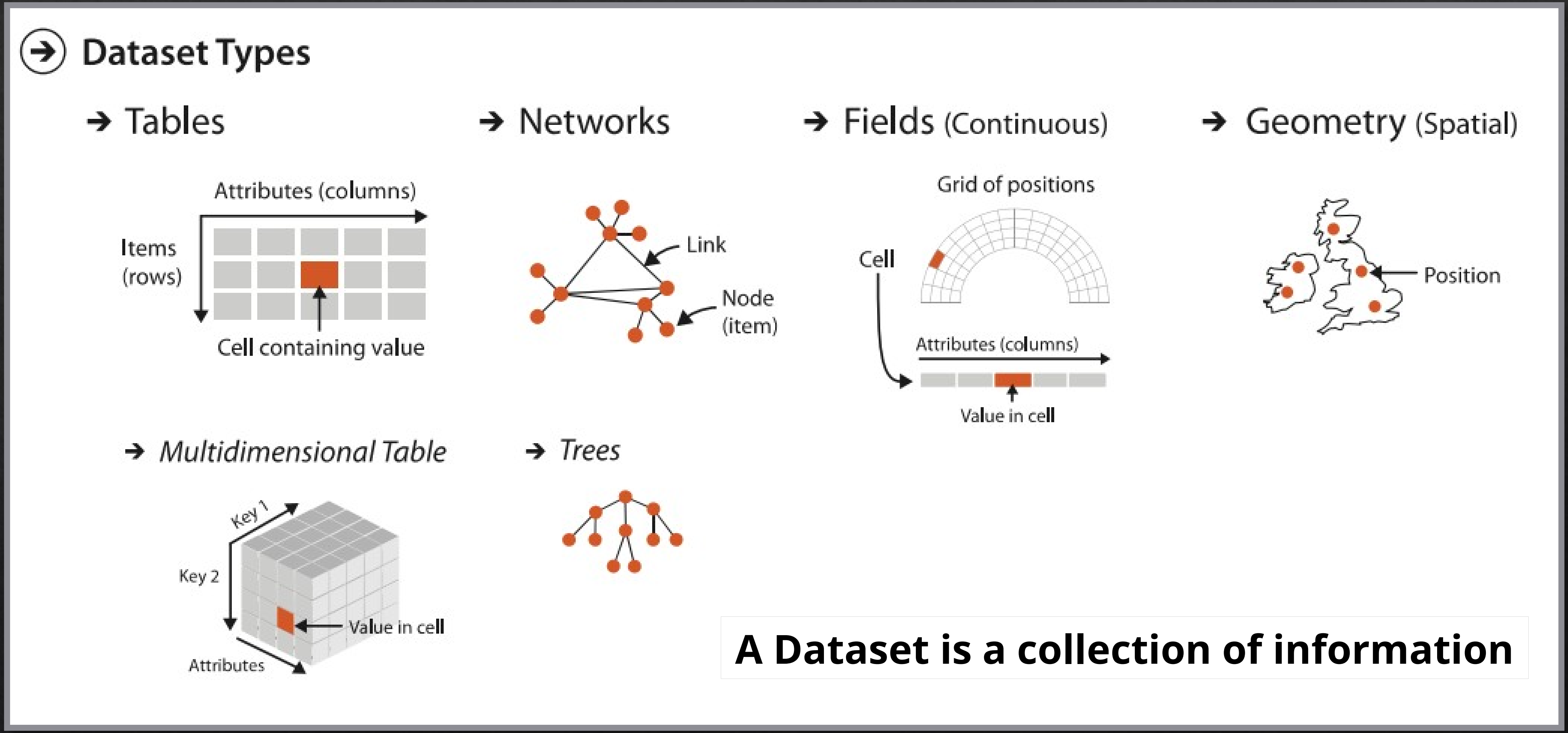
Datasets and Data Abstraction

Debajyoti Mondal

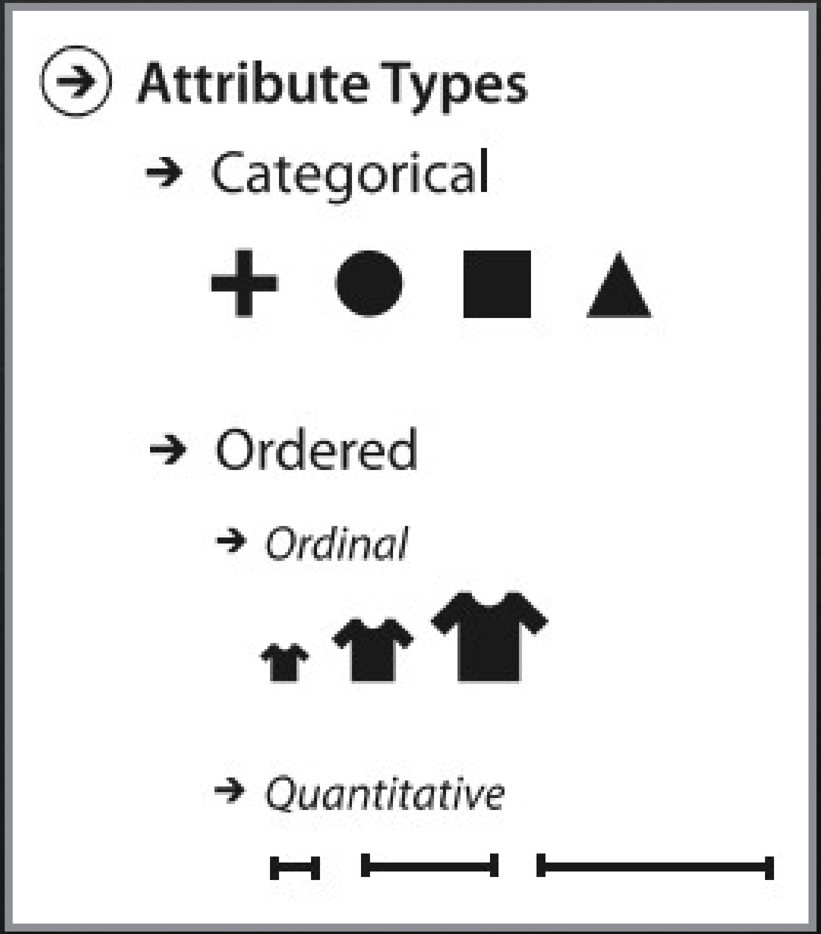
University of Saskatchewan

Example of Dataset Types



A Dataset is a collection of information

Other Properties of Data



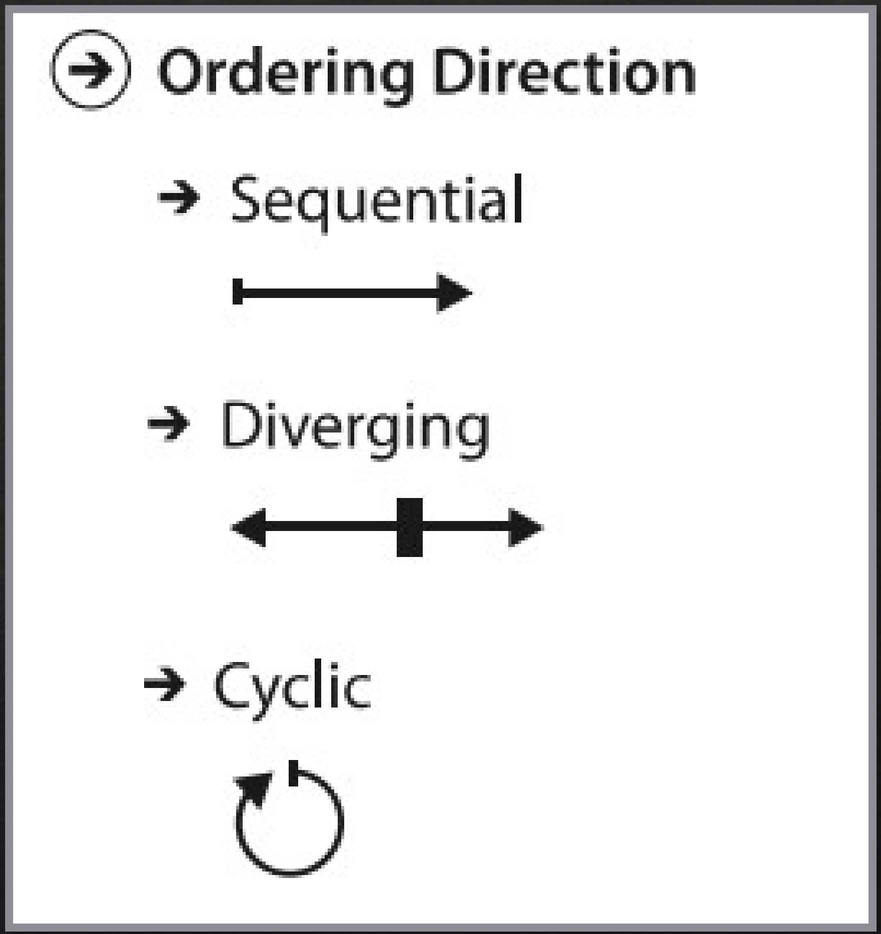
Attribute Types

Categorical

Ordered

Ordinal

Quantitative

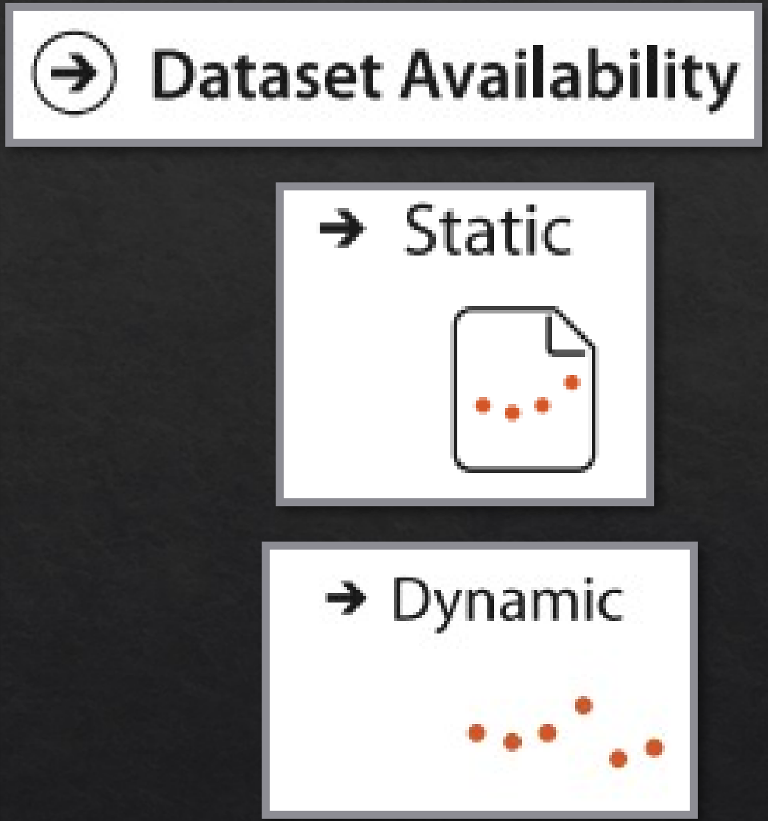


Ordering Direction

Sequential

Diverging

Cyclic



Dataset Availability

Static

Dynamic

Data Semantics & Type

Real-world meaning

Structural or Mathematical Interpretation

90210

Data Semantics & Type

Real-world meaning

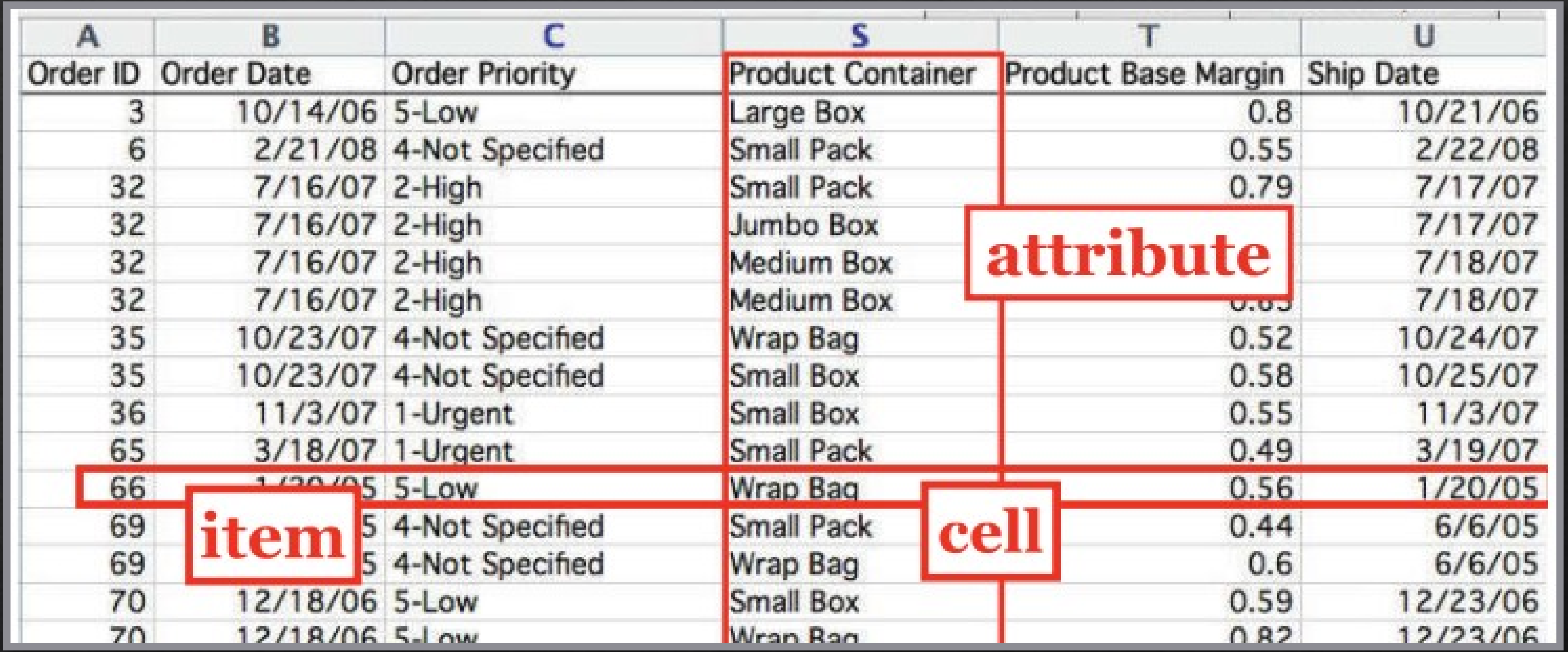
Structural or Mathematical Interpretation

90210

A postal code in CA □ Categorical

Number of Pizza Places □ Quantitative

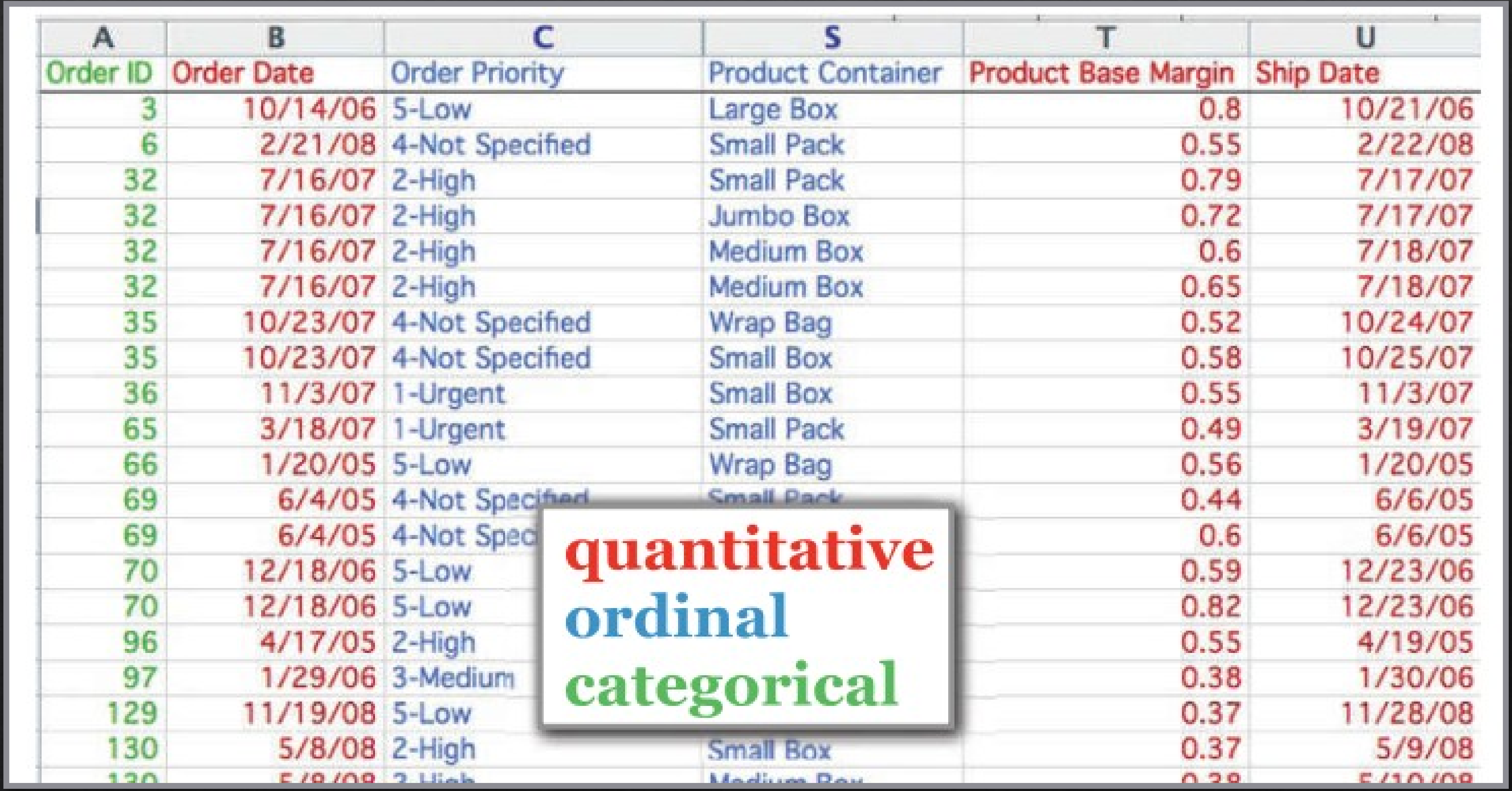
A Typical Table



A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	[illegible]	7/17/07
32	7/16/07	2-High	Medium Box	[illegible]	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	[illegible]	4-Not Specified	Small Pack	0.44	6/6/05
69	[illegible]	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06

Figure 2.5. In a simple table of orders, a row represents an *item*, a column represents an *attribute*, and their intersection is the *cell* containing the value for that pairwise combination.

A Typical Table



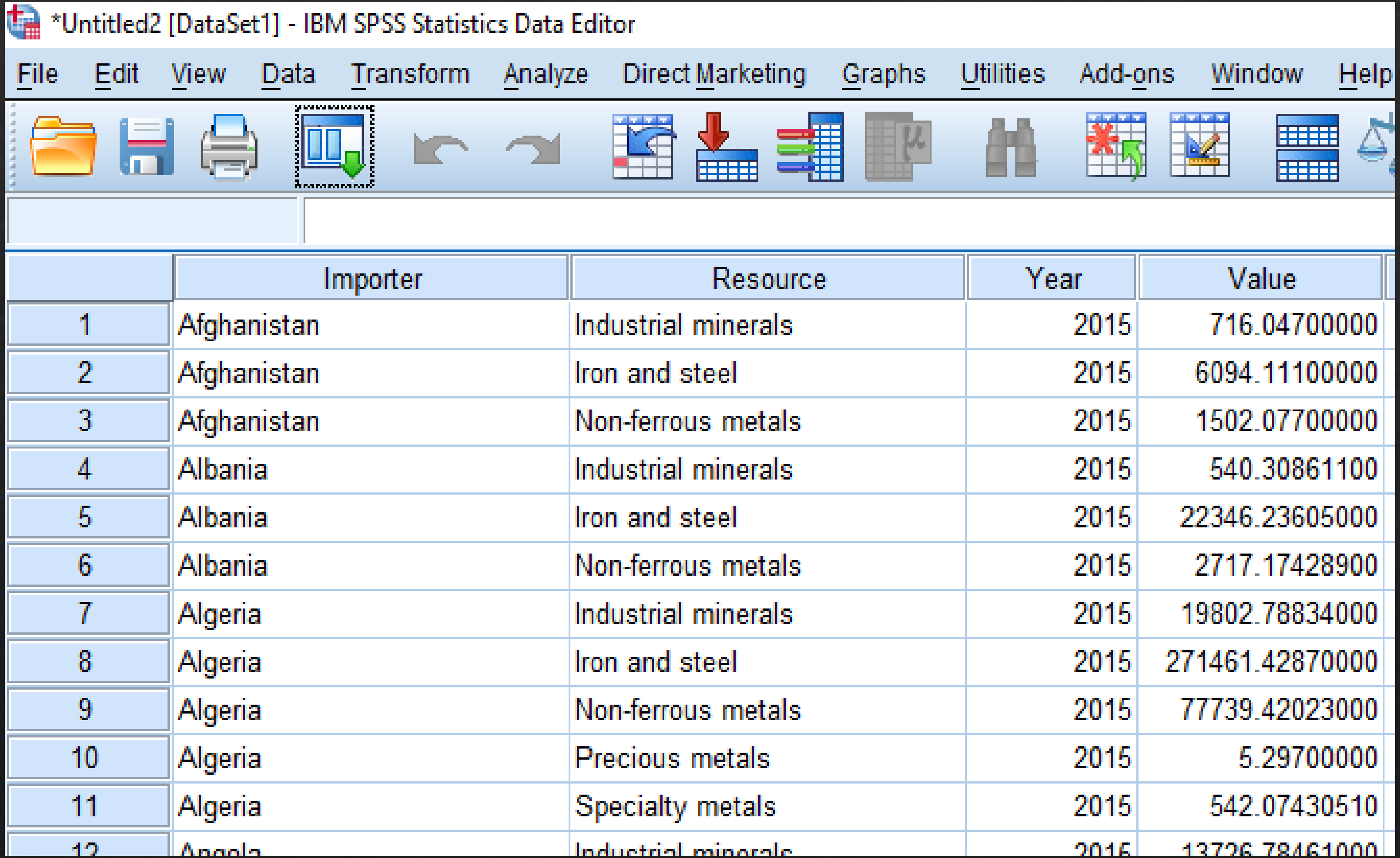
A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Spec		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08

quantitative

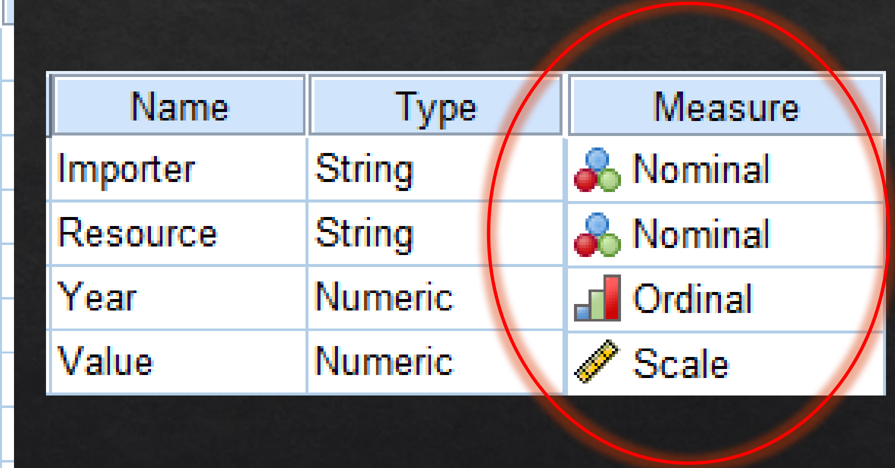
ordinal

categorical

Yes, we need this in practice
e.g., IBM SPSS Statistics

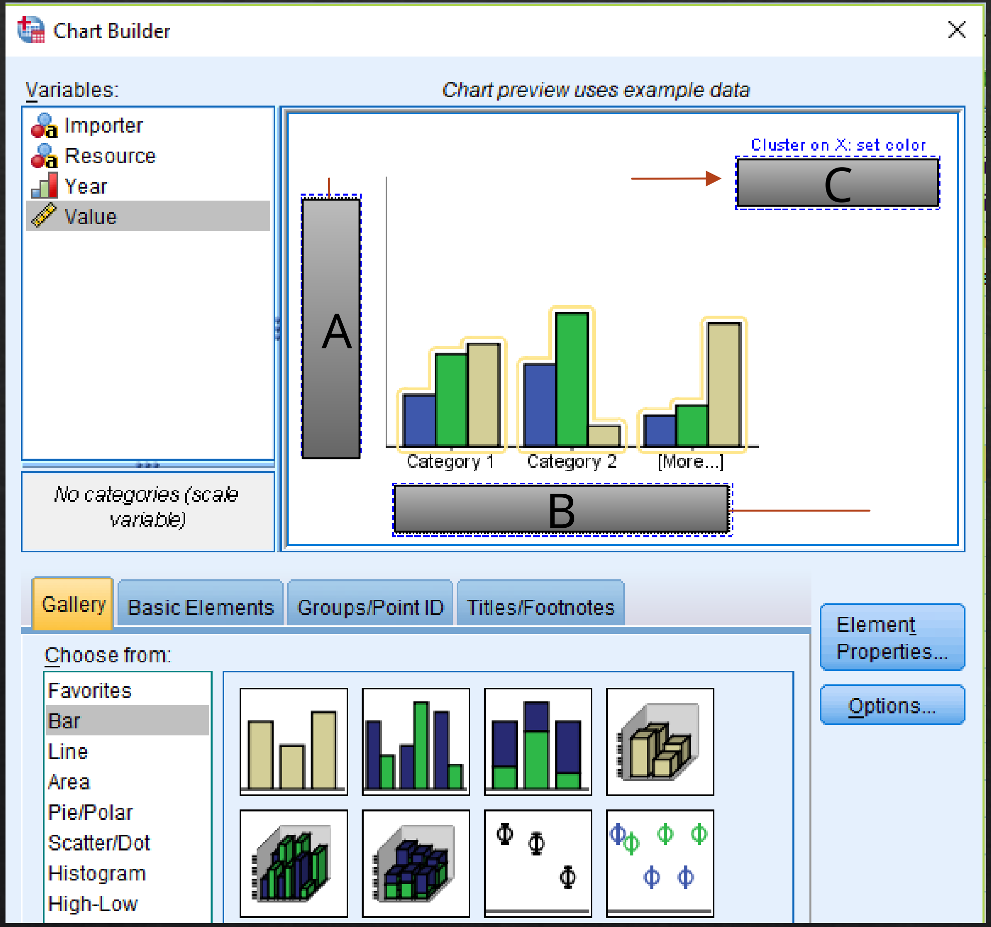


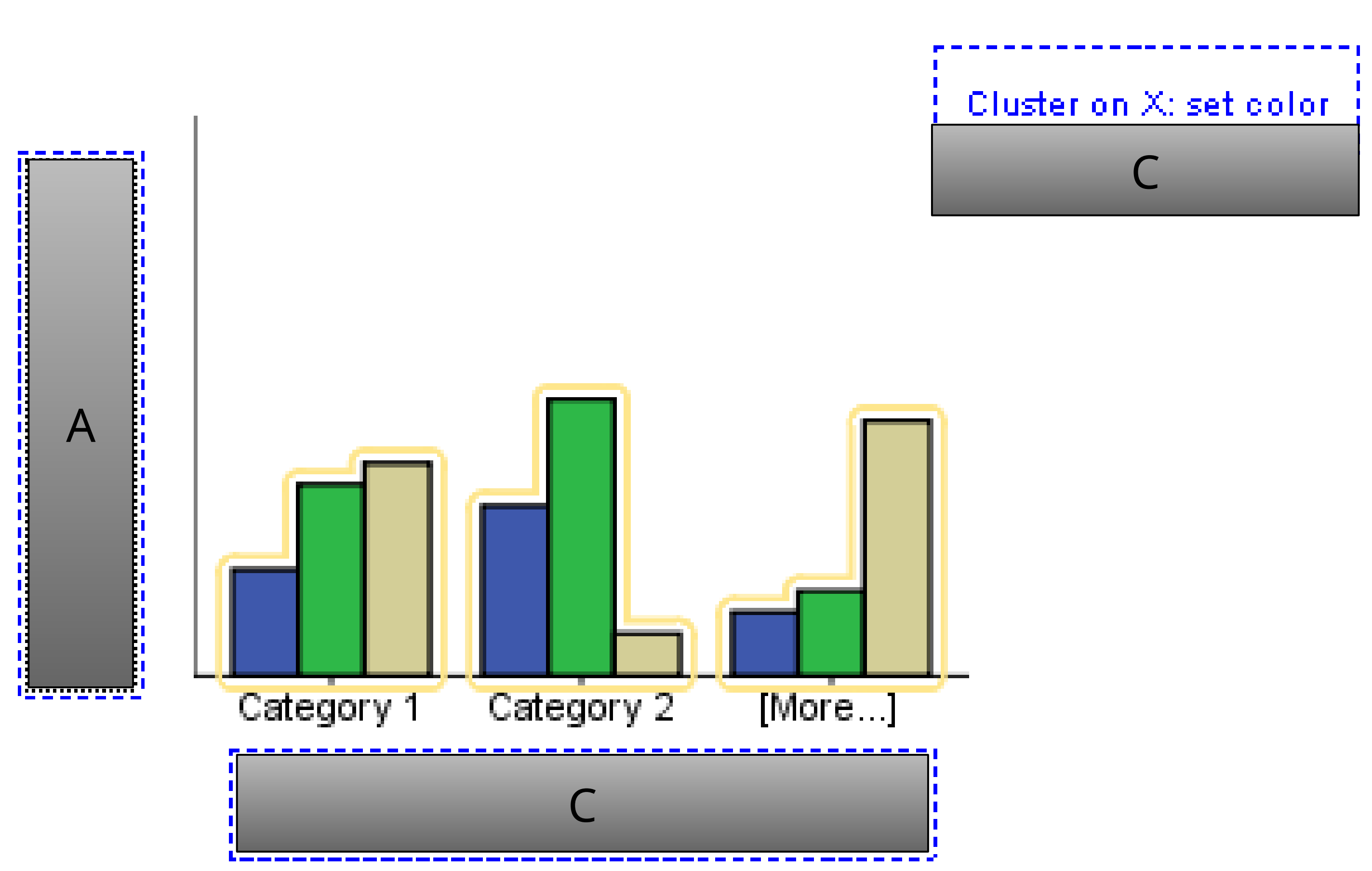
	Importer	Resource	Year	Value
1	Afghanistan	Industrial minerals	2015	716.04700000
2	Afghanistan	Iron and steel	2015	6094.11100000
3	Afghanistan	Non-ferrous metals	2015	1502.07700000
4	Albania	Industrial minerals	2015	540.30861100
5	Albania	Iron and steel	2015	22346.23605000
6	Albania	Non-ferrous metals	2015	2717.17428900
7	Algeria	Industrial minerals	2015	19802.78834000
8	Algeria	Iron and steel	2015	271461.42870000
9	Algeria	Non-ferrous metals	2015	77739.42023000
10	Algeria	Precious metals	2015	5.29700000
11	Algeria	Specialty metals	2015	542.07430510
12	Angola	Industrial minerals	2015	13726.78461000



Name	Type	Measure
Importer	String	Nominal
Resource	String	Nominal
Year	Numeric	Ordinal
Value	Numeric	Scale

IBM SPSS Statistics

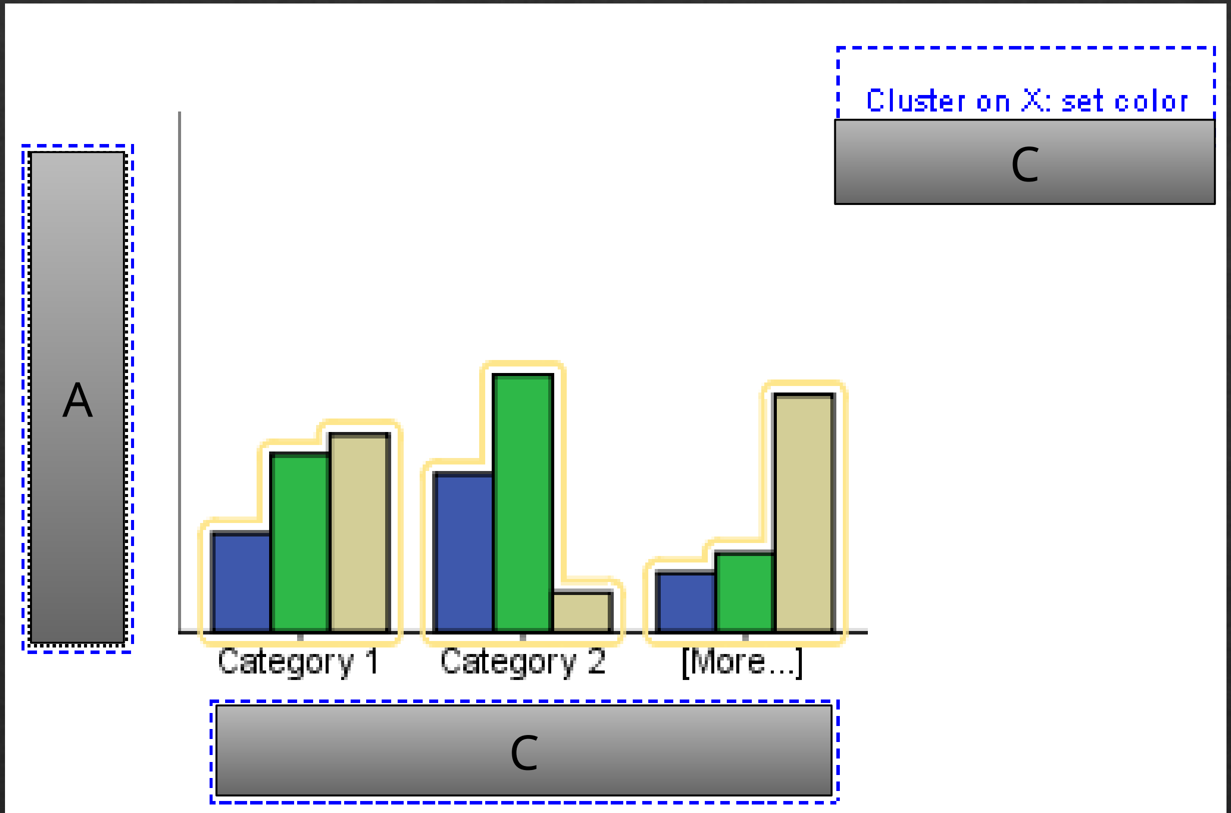


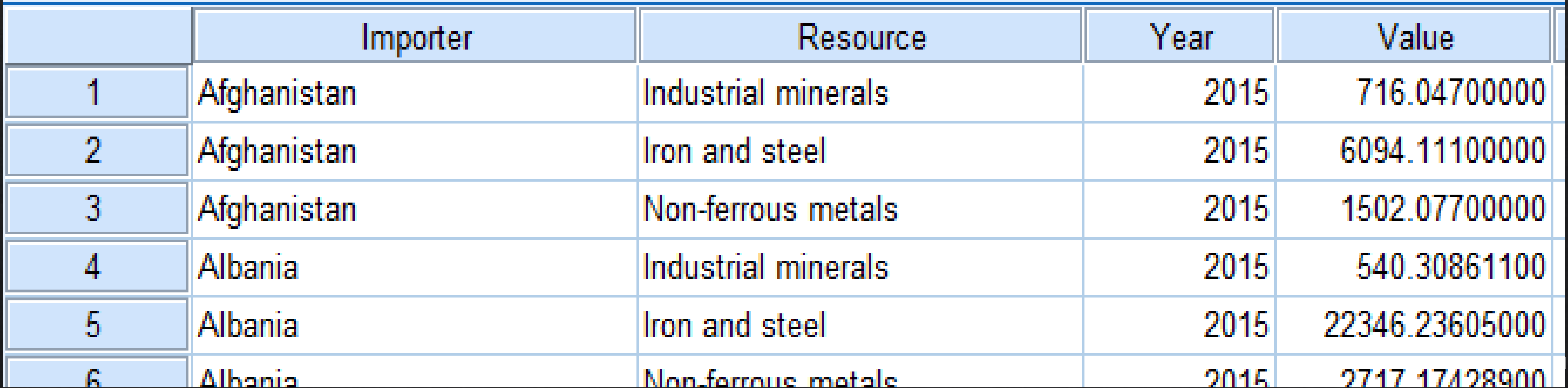


Class Activity

We want to examine the mean imported value for each resource.

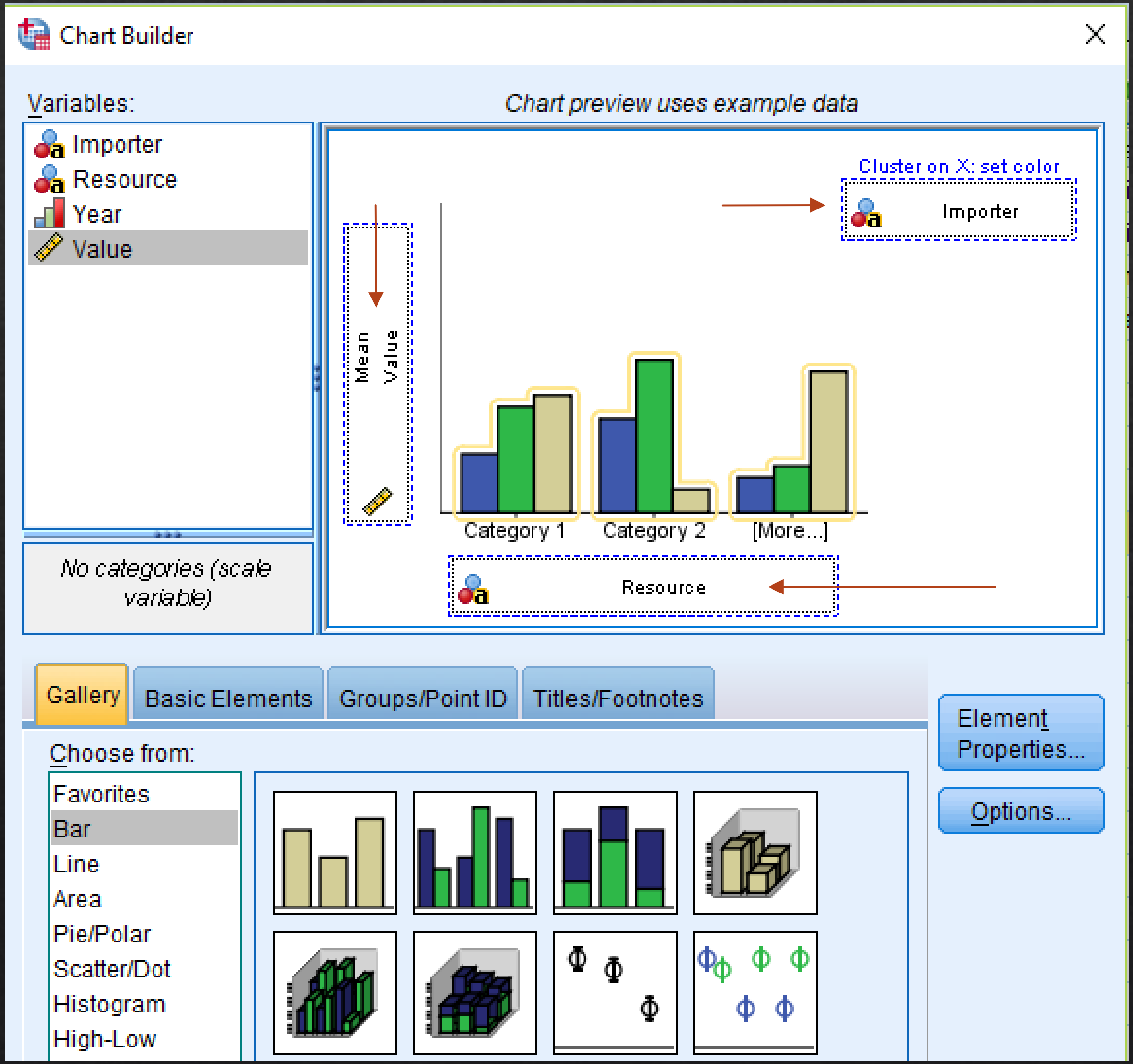
- What variables should we select for A, B, C?
- Draw a hypothetical final chart.



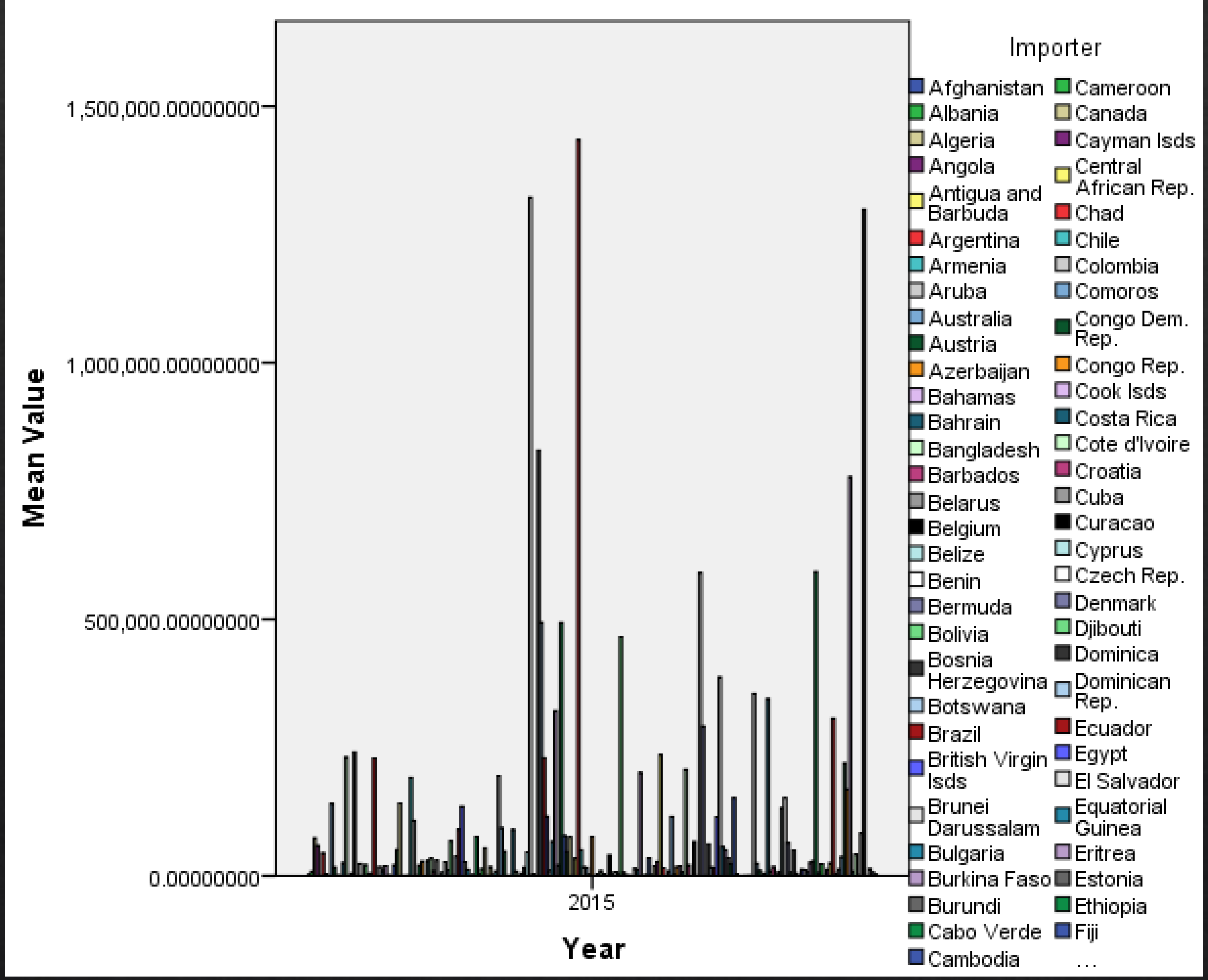


	Importer	Resource	Year	Value
1	Afghanistan	Industrial minerals	2015	716.04700000
2	Afghanistan	Iron and steel	2015	6094.11100000
3	Afghanistan	Non-ferrous metals	2015	1502.07700000
4	Albania	Industrial minerals	2015	540.30861100
5	Albania	Iron and steel	2015	22346.23605000
6	Albania	Non-ferrous metals	2015	2717.17428900

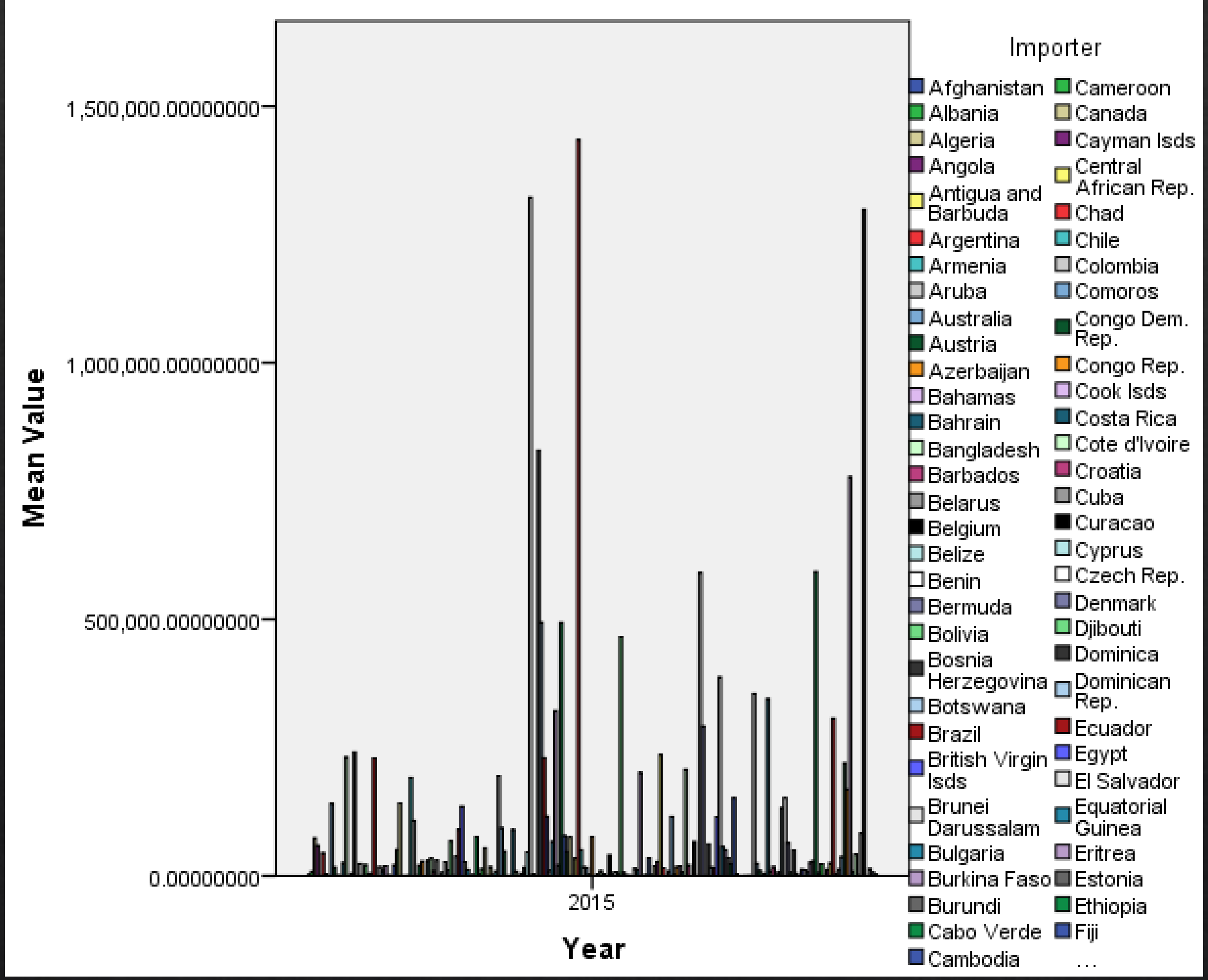
IBM SPSS Statistics

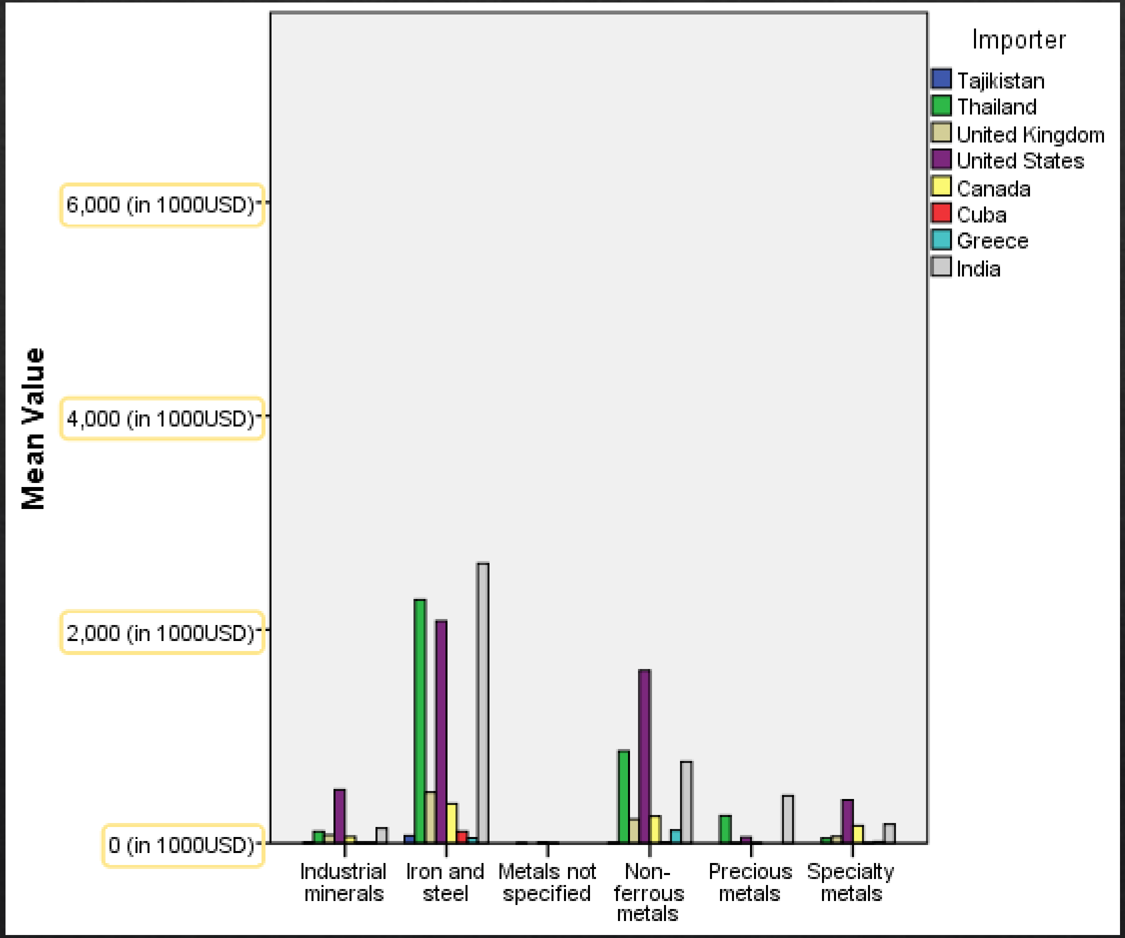


IBM SPSS Statistics



IBM SPSS Statistics


Mean Value
Year
Importer
Afghanistan
Albania
Algeria
Angola
Antigua and Barbuda
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bermuda
Bolivia
Bosnia Herzegovina
Botswana
Brazil
British Virgin Isds
Brunei Darussalam
Bulgaria
Burkina Faso
Burundi
Cabo Verde
Cambodia
Cameroon
Canada
Cayman Isds
Central African Rep.
Chad
Chile
Colombia
Comoros
Congo Dem. Rep.
Congo Rep.
Cook Isds
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Curacao
Cyprus
Czech Rep.
Denmark
Djibouti
Dominica
Dominican Rep.
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Ethiopia
Fiji


Mean Value
6,000 (in 1000USD)
4,000 (in 1000USD)
2,000 (in 1000USD)
0 (in 1000USD)
Industrial minerals
Iron and steel
Metals not specified
Non-ferrous metals
Precious metals
Specialty metals
Importer
Tajikistan
Thailand
United Kingdom
United States
Canada
Cuba
Greece
India