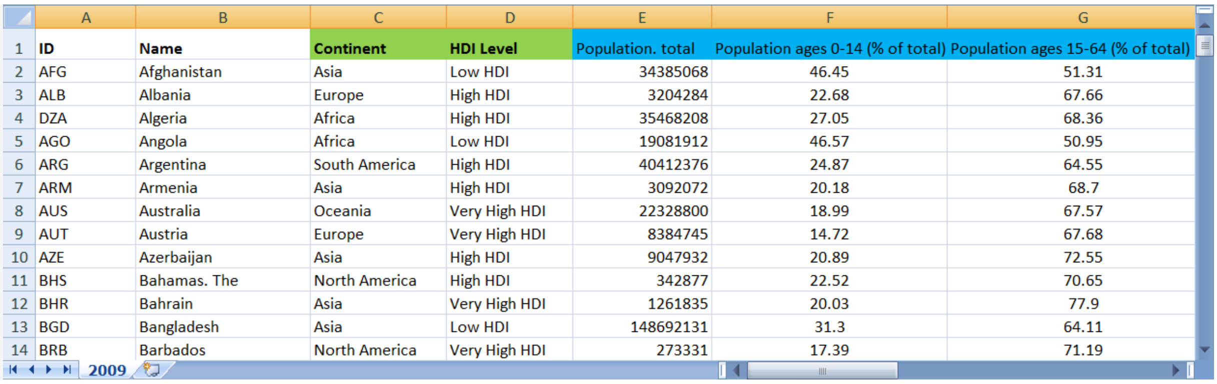
Introduction to Numerical and Categorical Data

Facilitated by the technological advances of the last decades, increasing amounts of complex temporal and multidimensional data are being collected from many different application fields such as business statistics, demographics, healthcare, biology, chemistry, energy, environment etc. A major challenge today is not to gather data, but to extract meaningful information and gain insights and knowledge from the data. Why is perception of multidimensional data a more difficult problem? An obvious answer is that we can barely perceive data values for two variables and as the amounts of gathered data and dimensions increases, the challenges become more complex and require innovative and interactive Visual Analytics tools.

A **data set** is defined as a collection of **data items**, where an item may, for instance, represent a country in a collection of statistical data. The characteristics of a data item are described by a collection of variables. A variable can be defined as a property, or a characteristic, of a data item that may vary from one item to another or over time. As an example below, the data items represent countries, the variables may represent various characteristics of the countries such as population size, percentage for various ageing groups or a category such as belong to a certain continent or a classified Human Development Index (HDI).

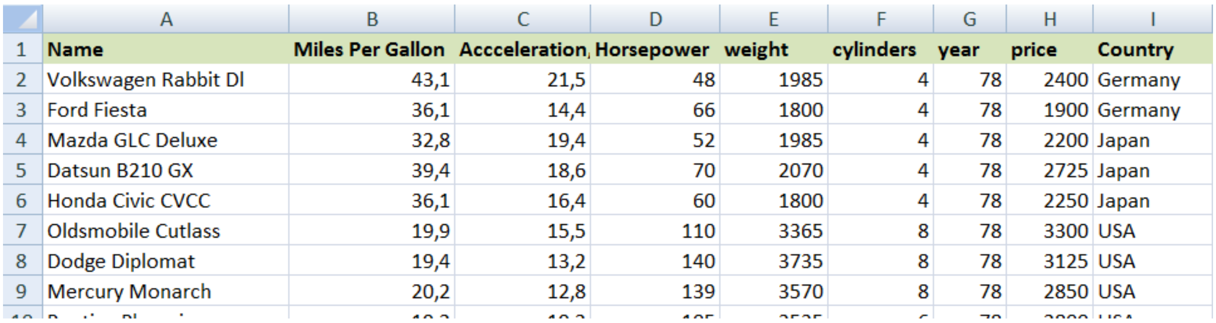
A **multivariate data set** is simply a data set including two or more variables. The items of a multivariate data set can be thought of as points in a multidimensional space where each dimension represents a variable. A standard format used for structuring multivariate data is to use an m-by-n matrix including m rows, usually representing data items, and n columns, usually representing variables. The table below displays a small example of a world data matrix where the first two columns represents the identification (ISO value and Name) and column C and beyond are data variables.



	A	B	C	D	E	F	G
1	**ID**	**Name**	**Continent**	**HDI Level**	Population. total	Population ages 0-14 (% of total)	Population ages 15-64 (% of total)
2	AFG	Afghanistan	Asia	Low HDI	34385068	46.45	51.31
3	ALB	Albania	Europe	High HDI	3204284	22.68	67.66
4	DZA	Algeria	Africa	High HDI	35468208	27.05	68.36
5	AGO	Angola	Africa	Low HDI	19081912	46.57	50.95
6	ARG	Argentina	South America	High HDI	40412376	24.87	64.55
7	ARM	Armenia	Asia	High HDI	3092072	20.18	68.7
8	AUS	Australia	Oceania	Very High HDI	22328800	18.99	67.57
9	AUT	Austria	Europe	Very High HDI	8384745	14.72	67.68
10	AZE	Azerbaijan	Asia	High HDI	9047932	20.89	72.55
11	BHS	Bahamas. The	North America	High HDI	342877	22.52	70.65
12	BHR	Bahrain	Asia	Very High HDI	1261835	20.03	77.9
13	BGD	Bangladesh	Asia	Low HDI	148692131	31.3	64.11
14	BRB	Barbados	North America	Very High HDI	273331	17.39	71.19

Data variables are here classified into two different types based on common classification taxonomy used in most visualization and data mining literature: **numerical** (quantitative) and **categorical** (qualitative). Categorical data can then be different names (Continent) or a classification that provide enough information to order the items (HDI Level).

The definition of variable types for a data set is important since various data types may require different visual analytic techniques since a single visualization technique is rarely appropriate for all types of data. Specifically, techniques used for numerical variables are often based on a numerical difference or similarity between data items. Categorical data on the other hand does not include any distance measure comparable to a numerical distance, and hence other visual analytical techniques may need to be used.



	A	B	C	D	E	F	G	H	I
1	**Name**	**Miles Per Gallon**	**Acceleration**	**Horsepower**	**weight**	**cylinders**	**year**	**price**	**Country**
2	Volkswagen Rabbit Dl	43,1	21,5	48	1985	4	78	2400	Germany
3	Ford Fiesta	36,1	14,4	66	1800	4	78	1900	Germany
4	Mazda GLC Deluxe	32,8	19,4	52	1985	4	78	2200	Japan
5	Datsun B210 GX	39,4	18,6	70	2070	4	78	2725	Japan
6	Honda Civic CVCC	36,1	16,4	60	1800	4	78	2250	Japan
7	Oldsmobile Cutlass	19,9	15,5	110	3365	8	78	3300	USA
8	Dodge Diplomat	19,4	13,2	140	3735	8	78	3125	USA
9	Mercury Monarch	20,2	12,8	139	3570	8	78	2850	USA

As a simple example based on a classic information visualization data set, the ranking of cars with “miles per gallon” may quite effectively be represented using our fish eye bar chart where the height of the bar represents the numerical variable consumption of gas, as displayed below. Using the same kind of representation for the categorical variable “Country” would, on the other hand, not be as useful since there does not exist any meaningful relationship between a country name and the height of a bar. However, we can use colour to show the similarity of two variables using colour for variable “Country”.

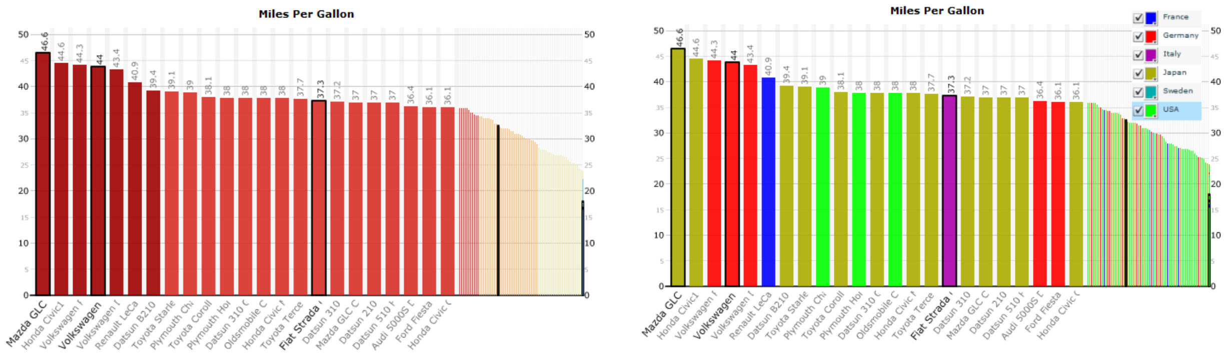


Figure: Classic car data set shown as bar chart for numerical variable “Miles per gallon” and coloured based on categorical variable Country.