CSE 332
Introduction to Visualization

Mini Project #1

Klaus Mueller

Computer Science Department
Stony Brook University

Rectangular Dataset

One data item

The variables
→ the attributes or properties we measured

The data items
→ the samples (observations) we obtained from the population of all instances

	A	B	C	D	E	F	G	H	I
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price
2	Volkswagen Rabbit Dl	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	Ford Fairmont A	USA	20,2	15,8	85	2965	6	78	2375
13	Ford Fairmont M	USA	25,1	15,4	88	2720	4	78	2275
14	Plymouth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100

Rectangular Dataset

Also called the *Data Matrix*

Car performance metrics

or Survey question responses

or Patient characteristics

....

One data item

Car models

or Survey respondents

or Patients

....

	A	B	C	D	E	F	
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylir
2	Volkswagen Rabbit DI	Germany	43,1	21,5	48	1985	
3	Ford Fiesta	Germany	36,1	14,4	66	1800	
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	
8	Dodge Diplomat	USA	19,4	13,2	140	3735	
9	Mercury Monarch	USA	20,2	12,8	139	3570	
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	
12	Ford Fairmont A	USA	20,2	15,8	85	2965	
13	Ford Fairmont M	USA	25,1	15,4	88	2720	
14	Plymouth Volare	USA	20,5	17,2	100	3430	
15	AMC Concord	USA	19,4	17,2	90	3210	
16	Buick Century	USA	20,6	15,8	105	3380	

Mission

Find some interesting data on the web

- something that challenges and interests you
- there are many data sources on the web
- use Google and some imagination (see also next slide)

Criteria for selection

- at least 250 data points (observations)
- at least 8 attributes
- the more the better (you can always reduce it)

Deliverables

- 2-3 page report that describes the data and justifies your choice
- see last slide for ore detail

Due date

- Friday, September 16, 11:59pm

Some Good Sources For Data

Kaggle – lots of data for data science

NYC Open Data – all kinds of data related to NYC operations

Kaiser Foundation – numerous data related to public health

Data.gov – open data site with US government data

Forbes – site with links to data sites

Data Quest – another site with links to data sites

Quandl – mostly financial and economics data

Open Data Inception – map w/data portals around the world

World Bank – collection of global development data

UCI repository – site that has been around for a long time

Analytics Vidhya – another site with many links to data sites

Wikipedia also has lots of data in tables

Dataset Example

Multivariate - Quantitative data and Categorical data

Data Items

	A	B	C	D	E	F	G	H	I
1	**Name**	**Country**	**Miles Per Gallon**	**Accceleration**	**Horsepower**	**weight**	**cylinders**	**year**	**price**
2	Volkswagen Rabbit Dl	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	Ford Fairmont A	USA	20,2	15,8	85	2965	6	78	2375
13	Ford Fairmont M	USA	25,1	15,4	88	2720	4	78	2275
14	Plymouth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100
25	Toyota Corona	Japan	27,5	14,2	95	2560	4	78	2975

Categorical

Quantitative

Categorical (Ordinal)
Quantitative

<u>Data types</u>

Quantitative (Numerical)
Categorical (Ordinal)

Notes on Dataset

Some advice

- avoid datasets where the majority of data is categorical (not overly exciting for binning, clustering, and so on)
- if an attribute is categorical it should have at least 6 categories
- convert textual categories into numbers by assigning a numerical ID
- aim for datasets with more than 250 data points and 8 attributes
- if your dataset is larger, pick 250 sample points at random (for now)
- if you have too many attributes keep the ones of interest (prefer quantitative attributes)
- if the data set has text, images, video, logs, etc. convert them to numbers via appropriate mechanism as discussed in class (this would be an advanced task, so you may want to avoid data like this)
- produce a spreadsheet of rows (data items) and attributes (columns)

Various Notes

Fusing data from two (or more) sources can yield interesting datasets

- you would have the same data points, just more attributes --> the attributes from both datasets
- goal: find two or more datasets that can show cause + effect for some theme

Examples:

- a dataset with crime data for all US States + datasets with state-wise average incomes, education levels, or unemployment rates might provide interesting insights on the causes of certain crimes
- a dataset with the weather per day at a set of cities + a dataset that has the on-time performances per day for the airports in these cities may provide insight on the reasons for the time delays

Example: Fusing Different Thematic Datasets

Address	Size	Bedrooms	Baths	Price	Zip Code
5 Nut Str.	2,345 sqft	3	1	$564k	11794

House listing data

Education by zip code

Zip Code	School Name	Avg. SAT	Class Size	Cost
11794	Tree Top	1060	34	Public

Quality of life by zip code

Zip Code	Livability Score	Distance to Airport	Air Quality Score	Electricity Cost
11794	63	45 miles	89	$0.34/KW

Make sure that all data are from the same/similar year (when time matters)

Might need different keys for linking different thematic datasets

- for example zip code, state, county, and so on
- find associations for each in all tables and fuse
- duplicate information for coarse grained tables in finer-grained tables

DELIVERABLES

The spreadsheet (one sheet) with all your data (10 points)

Explanation of each attribute (at least 8) (5 points/attribute, up to 10 attrib.)

- if numerical, how many values does it have, approximately
- if categorical, how many categories does it have (levels, like red, green, blue...)
- need to describe what the attribute is all about to earn the points

Information about the dataset(s)

- name the source(s) of the dataset(s) with URL (10 points)
- does it have at least 250 data points? (10 points)
- 10 extra points if it has 500 points or more
- 20 extra points if you fused two or more datasets together

Justification of your choice of data

- why are these data interesting ? (10 points)
- up to 5 hypotheses on what you might find and will prove/disprove by ways of visualization (5 points for each reasonable hypothesis, up to 5 hypotheses)