CSE 564
Visualization & Visual Analytics

Mini Project #1

Klaus Mueller

Computer Science Department
Stony Brook University

Purpose

Get a feel for data and where to find them

Get your hands dirty with JavaScript and D3.js

Online tutorials

- JavaScript: W3Schools.com
- D3: freeCodeCamp, D3 website, github makeBarChart ScottMurray exampleHub

Take advantage of this opportunity to learn D3 **now**

- you will need this later in the course

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	

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

Notes on Dataset

Some advice

- avoid datasets where the attributes only have a few different values
  - this applies to both categorical and numerical attributes
- convert textual categories into numbers by assigning a numerical ID
- if dataset is too large, reduce samples by random selection (for now)
- if you have too many attributes keep the ones of interest
- highly recommended: fuse multiple datasets together to get a more holistic analysis with a deep explanation of the ecosystem;
  - data on crime stats of a city plus data on education and data on demographics (3 files from different sources)
  - data on houses for sale (w/ house properties) plus data on the zip code of the houses such as education, crime, distance to airport, etc.
  - use Google to search for data such as “education quality by zip code”
- produce a spreadsheet of rows (data items), attributes (columns)
  - the goal is to add more columns (attributes)

Assignment (1)

Get some CSV-based data (see course slides for good sources)

- at least 250 data points (the more the better)
- at least 15 dimensions (fuse datasets to achieve this)
- good mix of numerical and categorical variables (minimum 5 levels)

Your D3-based visual interface should be able to (10 pts each):

1. present a menu to allow users to select a variable and update chart
2. draw a bar chart if a categorical variable is selected
3. draw a histogram if a numerical variable is selected (bin it into a fixed range (equi-width) of your choice)
4. on mouse-over display the value of the bar on top of the bar
5. on mouse-over also make the bar wider and higher to focus on it
6. mouse (with left mouse button down) move left (right) should decrease (increase) bin width/size (for numerical variables only)

Assignment (2)

An additional 10 pts for elegant implementation/function

Don't forget to

- label the axes (variable names)
- label the x-axis (bin range midpoints or category label)
- label the y-axis (number of items)

Due date is Tuesday, February 18 end of day

Submission on blackboard

Deliverables

Upload the following:

- 2-3 page report with illustrated description of your program's capabilities and implementation detail
- add code snippets to show how you did things
- YouTube link to a voice-narrated video file that shows all features of your software in action
- zip file with source code

Grading

- TA will pick students at random for thorough code review sessions
- you better know your code !!!
- so, please do not just copy code beyond the D3 templates
- or even worse, videotape someone else's program

Seeking Outside Help

Aka, cheating

Discussion with your class mates (but not others) is OK

Cut and paste from any source is not OK

- any suspected activity of this kind will result in zero points
- also for the person providing the original
- two-strikes and out rule is in effect (including an academic misconduct report)
- this includes any feeble attempt to cover the tracks somehow

Stay honest and resist the temptation!