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 dashingd3js tutorialsteacher

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
Data.World – lots of free datasets
Wikipedia also has lots of data in tables

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
  - some important variables may have fewer levels, such as gender (OK)
- convert textual categories into numbers by assigning a numerical ID
- aim for datasets with more than 500 data points and 8 attributes
- if your dataset is larger, pick 500 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 that can bring deeper and more comprehensive insight, broadening the 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

Assignment (1)

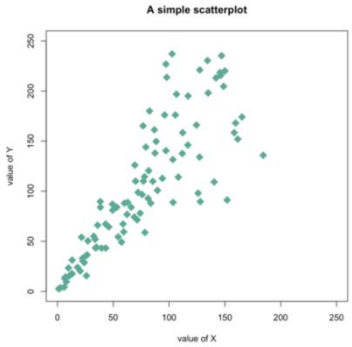
Get some CSV-based data (also see previous slides)

- at least 250 data points, but ideally at least 500
- at least 15 dimensions (fuse datasets to achieve this)
- good mix of numerical and categorical variables

Assignment (CS Students)

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 bar's value above the bar in red color
5. on mouse-over also color the bar in red to highlight it
6. mouse (with left mouse button down) move left (right) should decrease (increase) bin width/size (for numerical variables only)
7. produce a scatterplot of two selected variables (use a radio button to determine which of the two variable axes is to be loaded)



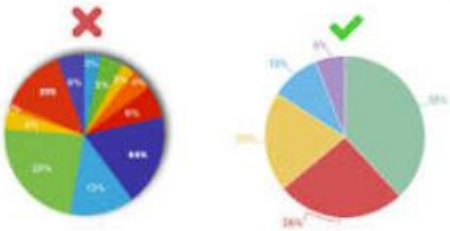
Assignment (non-CS Students)

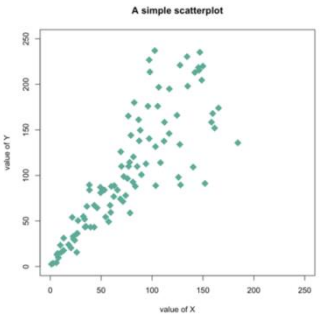
Choice of D3 is optional, else go with plotly (Dash)

- if you choose D3, then please use the CS-Student assignment

Your plotly-based visual interface should do (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. produce a pie chart of the selected variable (make sure you have no more than 5-6 slices, sum the smaller ones into a single 'others' slice)
5. produce a scatterplot of two selected variables (use a radio button to determine which of the two variable is to be loaded)





Assignment (3)

An additional 10 pts for overall elegant implementation and 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, March 2 end of day

Submission on blackboard

Deliverables

Upload the following to Blackboard:

- voice-narrated video file that shows all features of your software in action
- 2-3 page report
  - name the source(s) of the dataset(s) with URL
  - describe the attributes
  - write why you thought these data are interesting
  - mention anything noteworthy about implementation (beyond the video)
- zip file with all source code
- excel sheet with the fused data

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

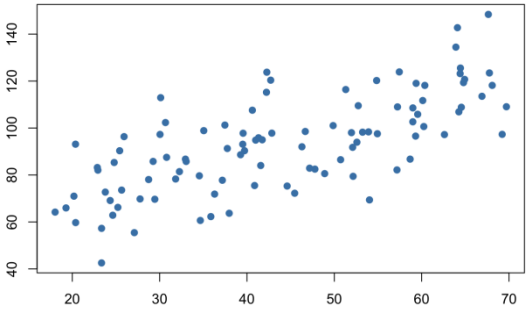
- any suspected activity of this kind will result in zero points
- also for the person providing the original
- two-strikes and out rule is in effect (including an academic misconduct report)
- this includes any feeble attempt to cover the tracks somehow

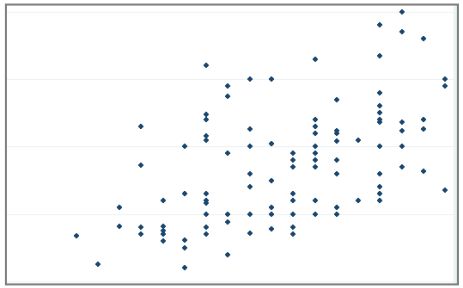
Stay honest and resist the temptation!

Notes on Scatterplots

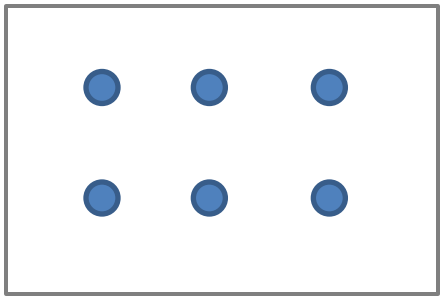
You can plot both numerical and categorical variables

- when one or both variables are numerical this is straightforward

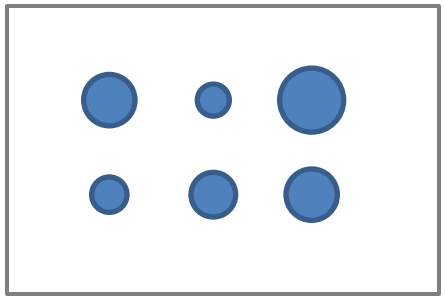




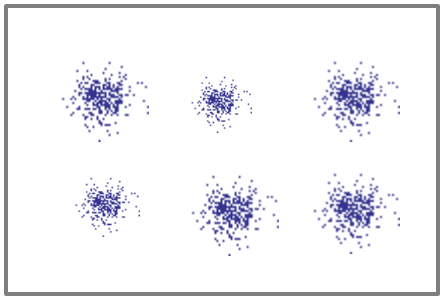
- when both variables are categorical you have three options



just overplot the points so they look like one point



scale disks in proportion to the number of points (more informative)



jitter each plotted point so a cluster with more points appears denser (more real)