47717: Data Mining
Homework #2: Clustering, Viz, Cleaning
Due: **21 May, 11:59pm, on Moodle**

Read the instructions **carefully** (that's a good idea in general).

- Each person submits their own theoretical part. The theoretical part should be a single file in pdf format only (no docx or jpg) named **ex2t_ID.pdf** (ID is your ID).
- If you are submitting handwritten answers, make sure they are crystal clear.
- Only **one** person from each group should submit the practical part code and answers. In the practical part submission link, the person who submits should enter the partner's name (using the add partner button). In addition, *both* people should write the name and id of their partners in the theoretical part pdf.
- For the practical part, the person who submits should submit a single ZIP file named **ex2p_ID.zip** (where ID is the ID of the person submitting). The zip should contain a folder named **code** and a folder called **output**.
- The meta question should be answered through a questionnaire. The link will be posted on the moodle.
- Points may be reduced for submissions that fail to comply.
- Make sure you follow the News forum for any updates.

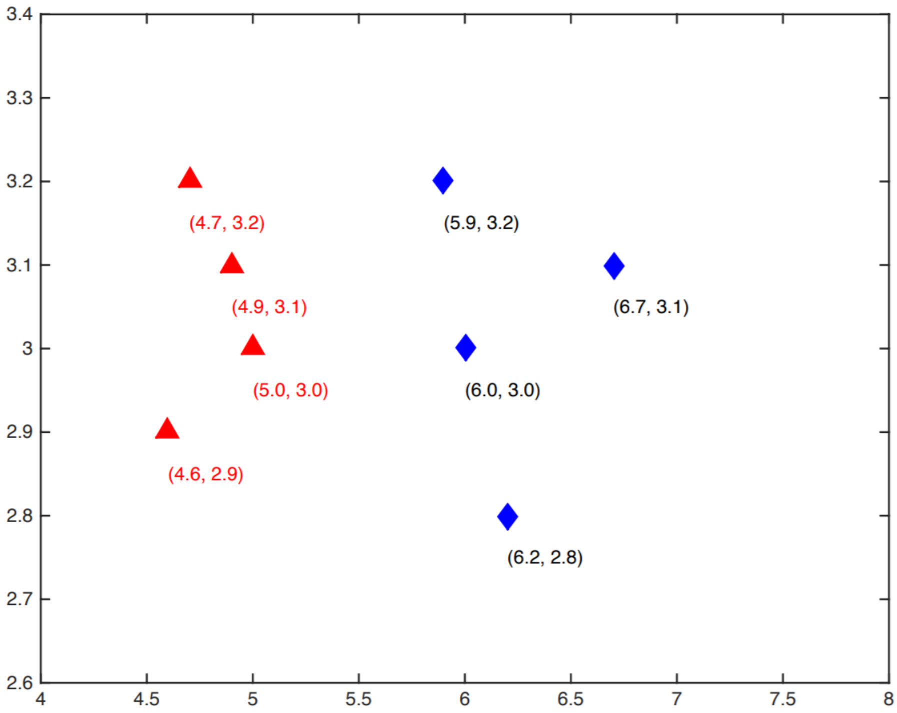
Problem 1 (Clustering).

(a) Consider the space of strings with edit distance as the distance measure. Give an example of a set of strings such that if we choose the clustroid by minimizing the sum of the distances to the other points we get one point as the clustroid, but if we choose the clustroid by minimizing the maximum distance to the other points, another point becomes the clustroid. Show the distances.

(b) Prove or provide a counterexample: k-means converges in a finite number of steps.

(c) k-means: Give an example of a dataset and a selection of k initial centroids such that when the points are reassigned to their nearest centroid at the end, at least one of the initial k points is reassigned to a different cluster.

(d) Hierarchical clustering:



In the figure above there are two clusters A (red) and B (blue), each has four members (coordinates in the figure). As we discussed, there are multiple ways to define distance between two (non-singleton) clusters. Compute the distance between A and B using the following methods. Use Euclidean distance for distance between members. Round your answer to four decimal places.

(1) Distance between the two farthest members? (called "complete link")
(2) Distance between the two closest members? ("single link")
(3) Average distance between all pairs? ("average")
(4) Among all three distances above, which are robust to noise?

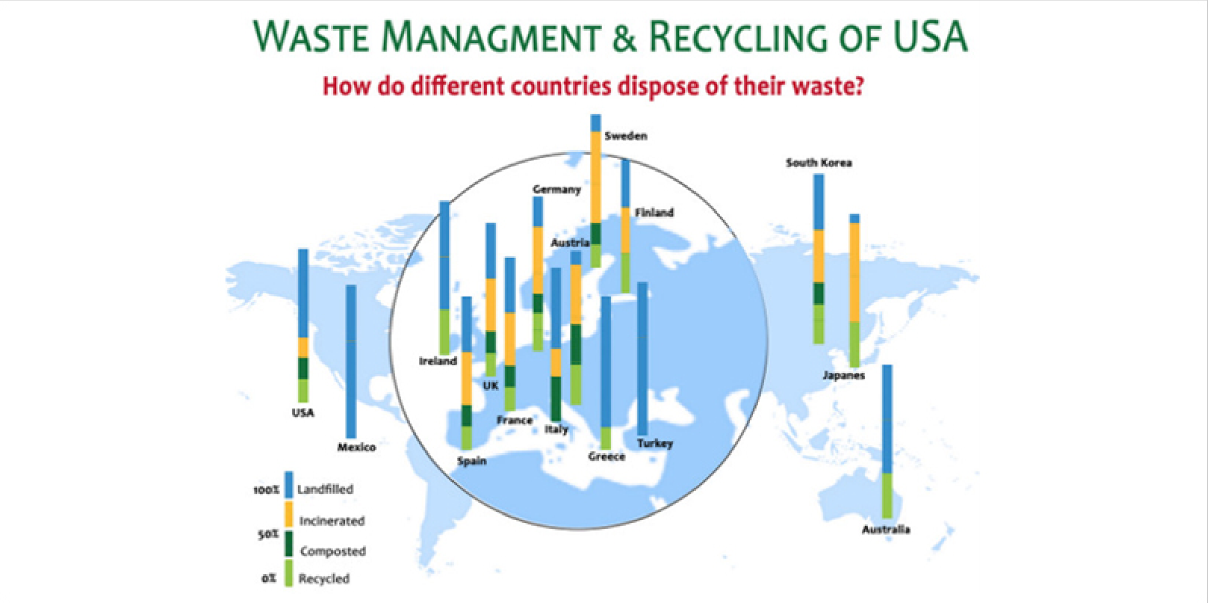
Problem 2 (Viz).

For each of the three figures below, and **one more bad visualization that you find** (and include in your answer):

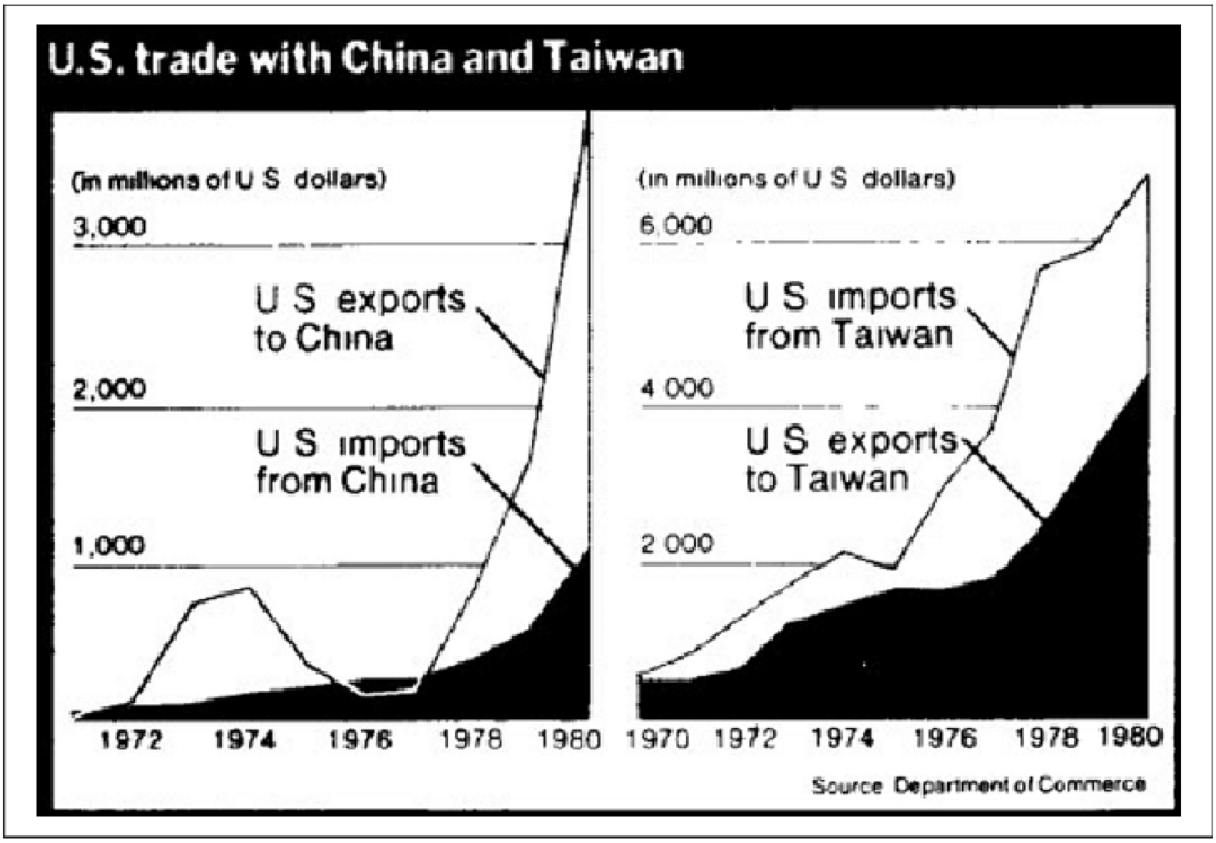
(a) Identify things that are wrong with the visualization.

(b) Explain what would be a better visualization and why (what can you infer from the new one you couldn't before?).

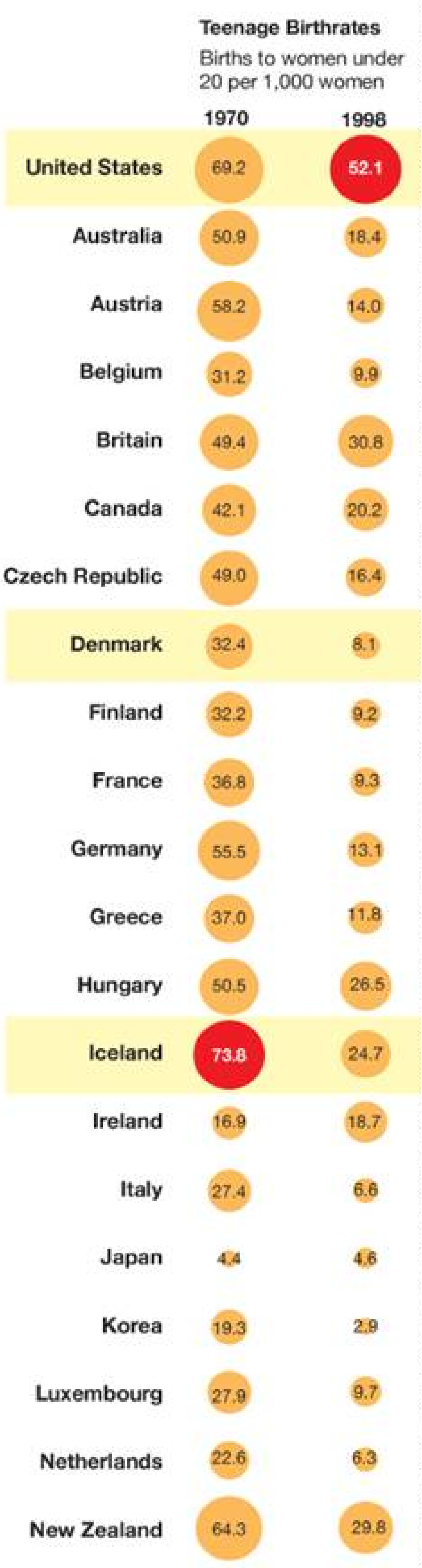
(c) Create a better one. You don't have to use the exact numbers from the figure if they are not stated (as in, I don't expect you to sit with a ruler and measure tiny bar charts) – it can be a sketch.


WASTE MANAGMENT & RECYCLING OF USA
How do different countries dispose of their waste?
Sweden
South Korea
Germany
Finland
Austria
Ireland
UK
France
Italy
Spain
Greece
Turkey
Japanes
USA
Mexico
Australia
100%
50%
0%
Landfilled
Incinerated
Composted
Recycled

(1)


U.S. trade with China and Taiwan
(in millions of U S dollars)
3,000
2,000
1,000
U S exports to China
U S imports from China
1972
1974
1976
1978
1980
(in millions of U S dollars)
6,000
4 000
2 000
U S imports from Taiwan
U S exports to Taiwan
1970
1972
1974
1976
1978
1980
Source Department of Commerce

(2)



Teenage Birthrates

Births to women under 20 per 1,000 women

	1970	1998
United States	69.2	52.1
Australia	50.9	18.4
Austria	58.2	14.0
Belgium	31.2	9.9
Britain	49.4	30.8
Canada	42.1	20.2
Czech Republic	49.0	16.4
Denmark	32.4	8.1
Finland	32.2	9.2
France	36.8	9.3
Germany	55.5	13.1
Greece	37.0	11.8
Hungary	50.5	26.5
Iceland	73.8	24.7
Ireland	16.9	18.7
Italy	27.4	6.6
Japan	4.4	4.6
Korea	19.3	2.9
Luxembourg	27.9	9.7
Netherlands	22.6	6.3
New Zealand	64.3	29.8

(3)

Problem 3 (Data Cleaning (Coding question)).

Remember I told you about the supermarket data? :) Here goes. Here's the background: `https://www.ynet.co.il/articles/0,7340,L-4658836,00.html`. The actual law is at `http://economy.gov.il/Trade/ConsumerProtection/Instructions/DocLib/O2015004355.pdf`. The files from a 1-day crawl are on the moodle.

(a) Find 7 things wrong with the data (7 different kind of issues, not the same issue in multiple supermarket chains). Can be anything from bad encoding and missing files to values that do not make sense, inconsistencies in product names, units... have fun with it). Show what you found. You don't need to fix it, unless it prevents you from finding more problems with the data (like bad encodings).

(b) Reality check: pick one field where you have some reasonable expectations about the data (price, zip code...) and create a histogram. Include it in your submission. Did it work? If not, what did you learn about the data?

Tip: The promotions table is somewhat confusing. I would start with items and stores.

Problem 4 (Meta). How long (in hours) did this assignment take? Please answer using the link, not in the pdf.