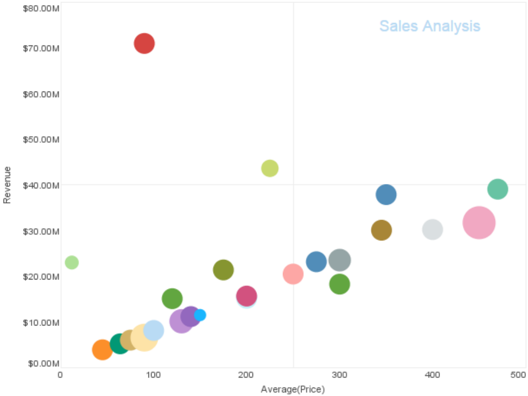
Data Visualization and Distances

Alipio Jorge - FCUP

October 2020

Visualization

- Communicate data clearly and effectively through graphical representation
  - Many dimensions?
    - how many dimensions can we show in a chart?
  - Visual accuity

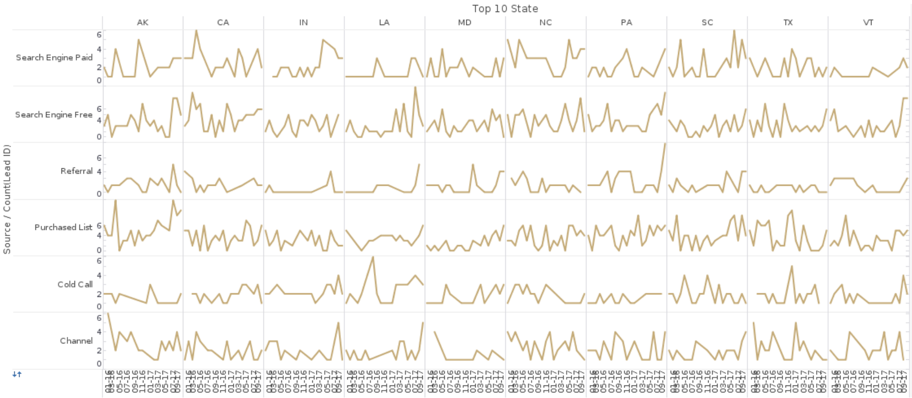


Visualization

- Overloading human ability to read compicated charts
  - We can see many dimensions
  - But **communication** may not be very effective

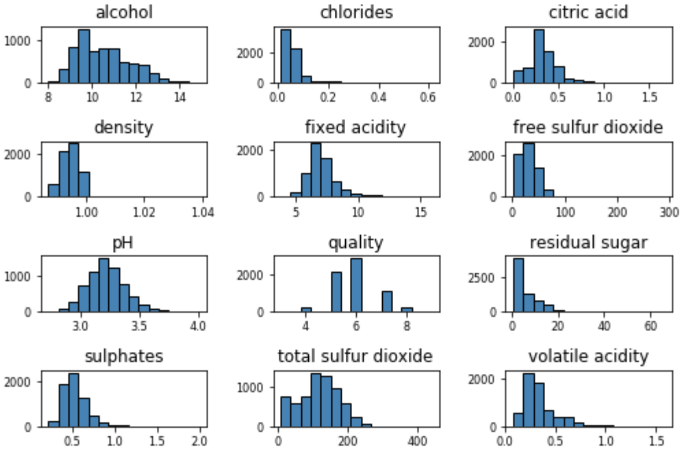
Visualization

- Effectiveness depends on what we want to show
  - Compare two lines in this chart: how does CA compare with TX?
  - Could we overlay all the lines for each variable?

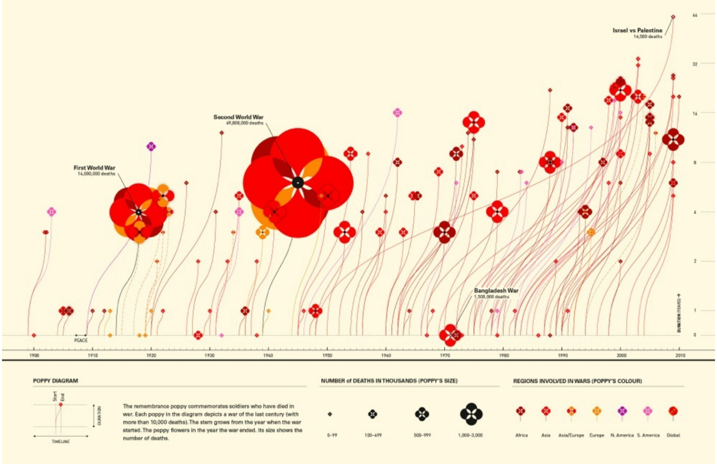


Visualization

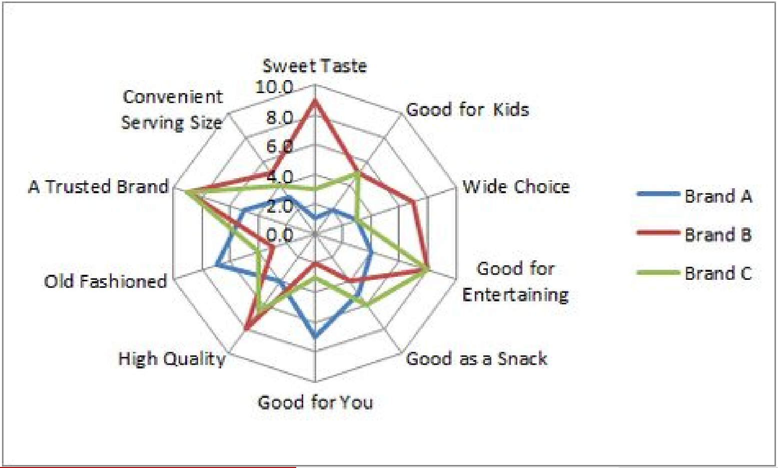
- But plain paneling is very useful
  - For data understanding
  - We are not comparing variables but looking at distributions



- Aesthetics is also important
  - but communication comes first
  - make your plot beautiful AND make sure it conveys the message



- Radar charts are strong at summarizing many descriptive dimensions
  - good for comparisons
  - can be arranged in a panel (to represent time for example)

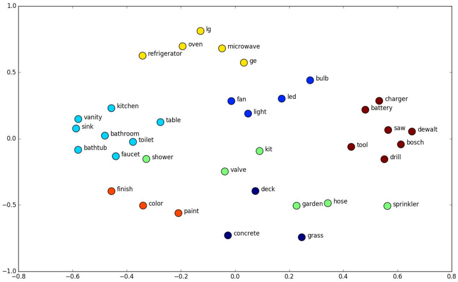


Visualization

- Worth also looking at
  - pie chart vs. bar chart (https://www.geckoboard.com/blog/pie-charts/)
  - animations
  - infographics
  - heatmaps

Data similarity and distance

- How similar are two data points?
  - Data vectors are points in a space
  - Similarity based approaches
    - **Applications**: Clustering customers, semantics of words with word embeddings, recommender systems
    - **Techniques**: Hierarchical clustering, k-means, k-nn, case based reasoning, collaborative filtering



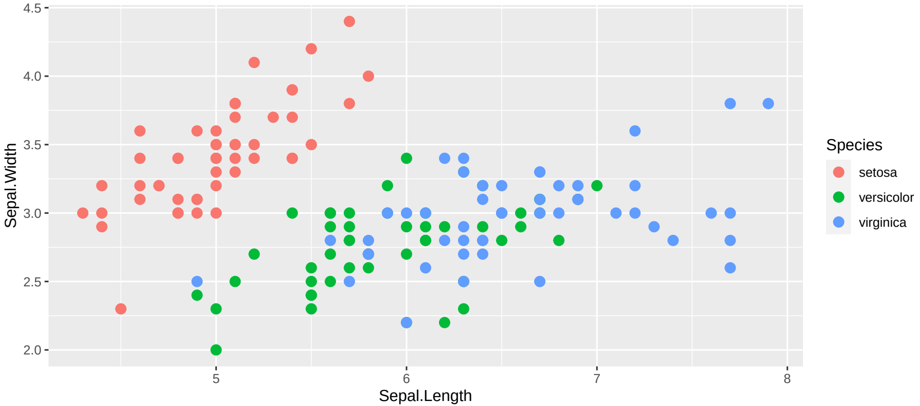
Data similarity and distance

- **Iris** data set: data points are 4d vectors + Species
  - 4 **measures of flower size** are dependent variables
  - **Species** is the independent variable

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                  
```

Data similarity and distance

- Closer points tend to belong to the same species
  - plot two of the variables coloured by species
  - How close are they?



Measuring distance

- Given two data points X_1 and X_2
  - measure the distance $d(X_1, X_2)$
  - Examples: think of two
    - iris **flowers** from the data set
    - **movies** according to length, genre, director, year, cast

Measuring distance

- Given two data points X_1 and X_2
  - measure the distance $d(X_1, X_2)$
  - many different **measures**
  - depending on the data types
  - Similarity can be computed from distance
    - if $0 \geq d(.,.) \leq 1$
    - $sim(.,.) = 1 - d(.,.)$

Numeric data

- Numeric data
  - Manhattan distance
  - Euclidean distance
  - Minkowski distance
  - Supremum or Chebyshev distance

Most comon measures for numeric data

- Numeric data
  - Manhattan distance

$$d(X, Y) = \sum_i |X_i - Y_i|$$

- Euclidean distance

$$d(X, Y) = \sqrt{\sum_i (X_i - Y_i)^2}$$

Generalizing Manhattan and Euclidean

- Numeric data
    - Minkowski distance is a generalization of Manhattan and Euclidean
    - if $p = 1$ we have manhattan, with $p = 2$ we have Euclidean

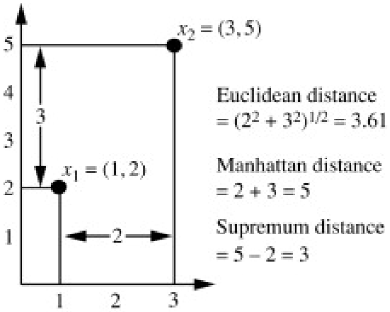
$$d(X, Y) = \sqrt[p]{\sum_i (X_i - Y_i)^p}, \quad h \geq 1$$

Norms

- Some notation
  - Euclidean distance is L_2 **norm** also $\|X\|_2$
  - Manhattan's is L_1 norm also $\|X\|_1$
  - Minkowski's is L_p norm, or $\|X\|_p$

Supremum distance

- What is L_∞ ?
  - Chebyshev or Supremum distance
  - corresponds to the **maximum coordinate distance**, $\max_i |X_i - Y_i|$



What is a distance measure?

- Not all measures $d(.,.)$ are distances, though...
  - A distance must have the following properties
    - Non-negative: $d(.,.) \geq 0$
    - Identity: $d(X, X) = 0$
    - Symmetry: $d(X, Y) = d(Y, X)$
    - Triangle inequality: $d(X, Z) \leq d(X, Y) + d(Y, Z)$

Nominal attributes

- Nominal or binary attribute
  - match (distance 0) or no match (distance 1)
  - Example: same director
- Nominal vector
  - $X_1 =< Comedy, \ TerryJones, \ Color, \ UK >$
  - $X_2 =< \ Scifi, \ TerryGilliam, \ Color, \ UK >$
  - $d(X_1, X_2) = 2/4 = 0.5$

$$d_{symetric}(X, Y) = \frac{\#features - \#matches}{\#features}$$

Jaccard similarity

- Binary vectors
  - Compare two movies M and N and by cast
  - We can use **binary** vectors
    - the dimensions $Actor_i$ describe actor participation
    - $M_i = 1$ if actor $Actor_i$ participated in the film
  - These vectors have mostly zeros (think of all the actors)
  - Symetric distance does not work (very small values)
  - Binary or Jaccard **similarity** (distance is $1 - sim$)

$$sim_{Jacc} = \frac{\#(Actors_M \cap Actors_N)}{\#(Actors_M \cup Actors_N)}$$

The distance / similarity matrix

- From the data matrix $[D_{i,j}]$
  - we obtain the **distance matrix** $[d(D_{i,.}, Dj,.)]$
  - or the **similarity matrix** $[sim(D_{i,.}, Dj,.)]$

$$S = \begin{bmatrix} 1 & 0.2 & 0.7 \\ 0.2 & 1 & 0 \\ 0.7 & 0 & 1 \end{bmatrix}$$

- This is quite handy for
  - e.g., clustering, similariry based classification

Data similarity and distance

- More things to consider
    - Hybrid descriptors (of varied types)
    - Cosine similarity
    - Pearson correlation
    - Ordinal attributes
    - Weighted distances / similarities

When objects are described with different types

- Describe movies with $< Year, \; Diretor, \; Genre, \; Length, \; Cast >$
- $M =< 1979, \; TerryJones, \; Comedy, 94, < JohnCleese, MichaelPalin, ... >>$
- $N =< 1983, \; TerryJones, \; Comedy, 107, < JohnCleese, TerryGilliam, ... >>$

When objects are described with different types

- How to compute similarity between M and N ?
  - All attributes should have the same scale
  - We find attribute-wise similarities in the $[0, 1]$ interval
    - $sim_{year}(y_1, y_2) = 1 - |y_1 - y_2|/150$ (the bug of the year 2042?)
    - $sim_{dir} = 1$ if the director is the same and 0 otherwise
    - $sim_{cast} = sim_{Jacc}$

$$sim(M, N) = \frac{1}{5} \sum_i sim(M_i, N_i)$$

When objects are described with different types

- We can give different weight to different attributes

$$sim_w(M, N) = \frac{1}{5} \sum_i w_i \; sim(M_i, N_i)$$

Cosine Similarity

- **Cosine similarity** is used when all the variables are in the same scale and are numerical
  - e.g. finding amazon users who are similar to me and get **good recommendations**
  - $U =< r_1, r_2, ..., r_n >$, where r_i are product ratings (5 stars)
  - e.g., $U =< 3, 2.5, 0, 0, 0, 5, 0, 1 >$

$$sim(U, V) = cos(U, V) = \frac{U.V}{\|U\|.\|V\|}$$

- Exercise: if U and V are binary *cos* can be calculated using set operations (intersection and length). How?

References and sources

- Data Mining Concepts and Techniques, Han, Kamber & Pei, Morgan Kaufmann
- Some charts obtained from https://www.inetsoft.com/blog/multidimensional-charting-many-dimensions-many/
- Good examples in pandas: https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57