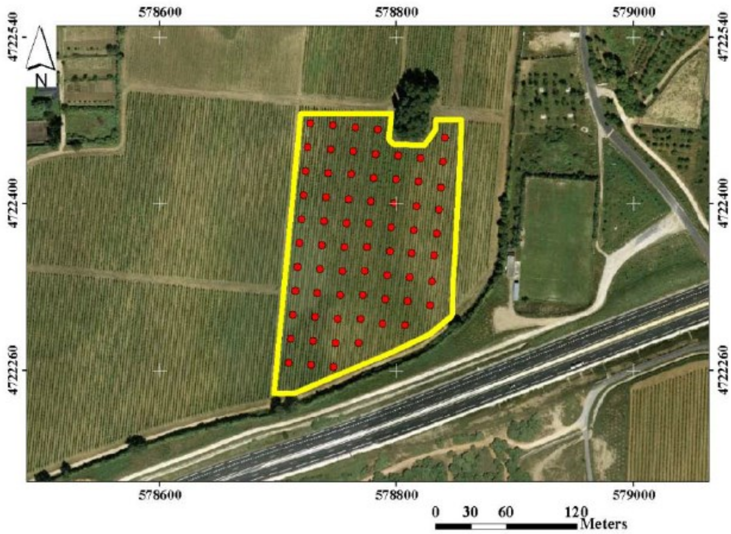
Topic 5

Sampling

Sampling

- **Randomly**: it can be done by randomly picking cases from the original dataset
- **Systematic**: It is possible to reduce the data using a sampling strategy that preserves the distribution.
  - It can be done by simple selecting regularly spaced data (systematic sampling). Example: examine every 100th item; divide an area by equal sized grids, pick a random grid and then every 5th grid is inspected.
  - However, it can result on information loss ("maps with holes")
- **Cluster**: the dataset is divided in groups and some groups are selected randomly.
- **Convenience**: choose the easiest to access.
  - The sample is not representative of the population. Therefore, it does not allow generalizations of results. However, it can be good for exploratory analysis.

Regular Sampling Grid



This is "systematic sampling" strategy

33

Sampling

- Another strategy involves the average on a neighborhood or a random selection on a certain region
  - There are many ways to do **random number generator in R** to create samples
    - Mainly according to the type of distribution that is wanted
    - Examples are:
      - `runif(n,min,max)    # random-uniform`
      - `rnorm(n,mean,std)   # random normal-gaussian`
    - Other examples
      `rbinom, rpois, rexp, rgamma, rlogis, rt, rchisq`, etc.

34

Topic 6

Dimensionality reduction

Dimensionality Reduction

- Sometimes it is necessary to reduce the data dimensionality so we can use certain visualization techniques
- This reduction should **preserve**, as much as possible, the **information** contained on the original data

Motivation:

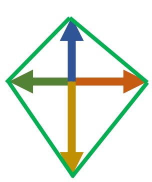
- **Curse of Dimensionality**: As the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse. This sparsity is problematic methods that requires statistical significance. This leads to models that overfit the training data and therefore perform poorly on unseen data.
- **Noise Reduction**: High-dimensional data tend to have more noise. By reducing the dimensionality, we can eliminate irrelevant features and reduce noise.
- **Improved Performance**: High-dimensional datasets are often computational and resource demanding. Reducing dimensionality can lead to less computational requirements.
- **Visualizing Data**: When dealing with a 2D or 3D dataset, it is possible to visualize the entire dataset. However, for data that has more than three dimensions, we need to use dimensionality reduction to project it into a 2D or 3D space to visualize it.
- **Avoiding Multicollinearity**: In high dimensions, variables may be highly correlated. Highly correlated variables do not provide unique information for model learning, leading to instability in coefficient estimates.

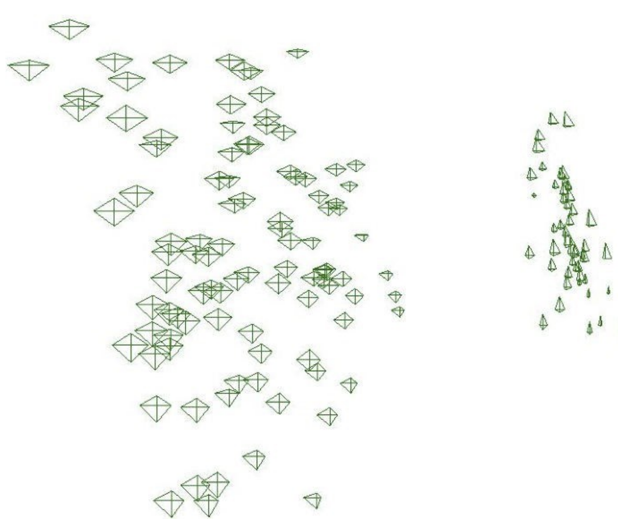
Dimensionality Reduction

- Such reduction can be made by hand, selecting attributes, or by using some established technique. Examples:
  - Principal Component Analysis (PCA) *
  - Multidimensional Scaling (MDS)
  - Self-Organizing Maps (SOM)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
  - Linear Discriminant Analysis (LDA) [supervised method]
  - Auto Encoders…

Dimensionality Reduction by the marker

PCA of the Iris dataset.
The glyphs represent the 4 original variables: each line from the center is proportional to an attribute values.





Let's see one project about Social Media publication strategies:

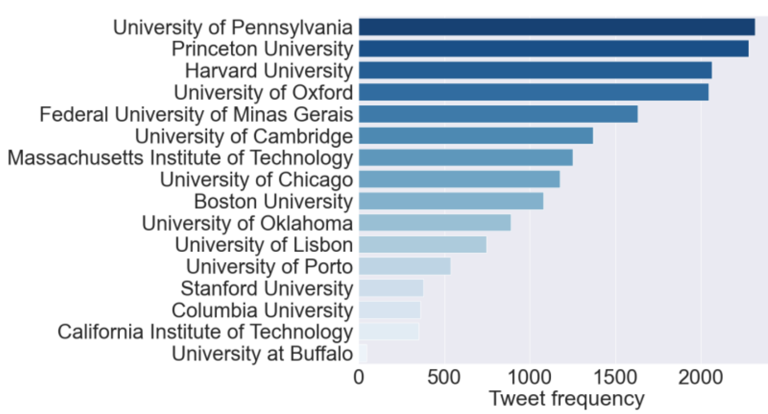
University of Pennsylvania
Princeton University
Harvard University
University of Oxford
Federal University of Minas Gerais
University of Cambridge
Massachusetts Institute of Technology
University of Chicago
Boston University
University of Oklahoma
University of Lisbon
University of Porto
Stanford University
Columbia University
California Institute of Technology
University at Buffalo

HEIs

Álvaro Figueira • VD • 2023 • 1ª ed.

39

Let's see one project about Social Media publication strategies:



HEIs

Variables: → for clustering

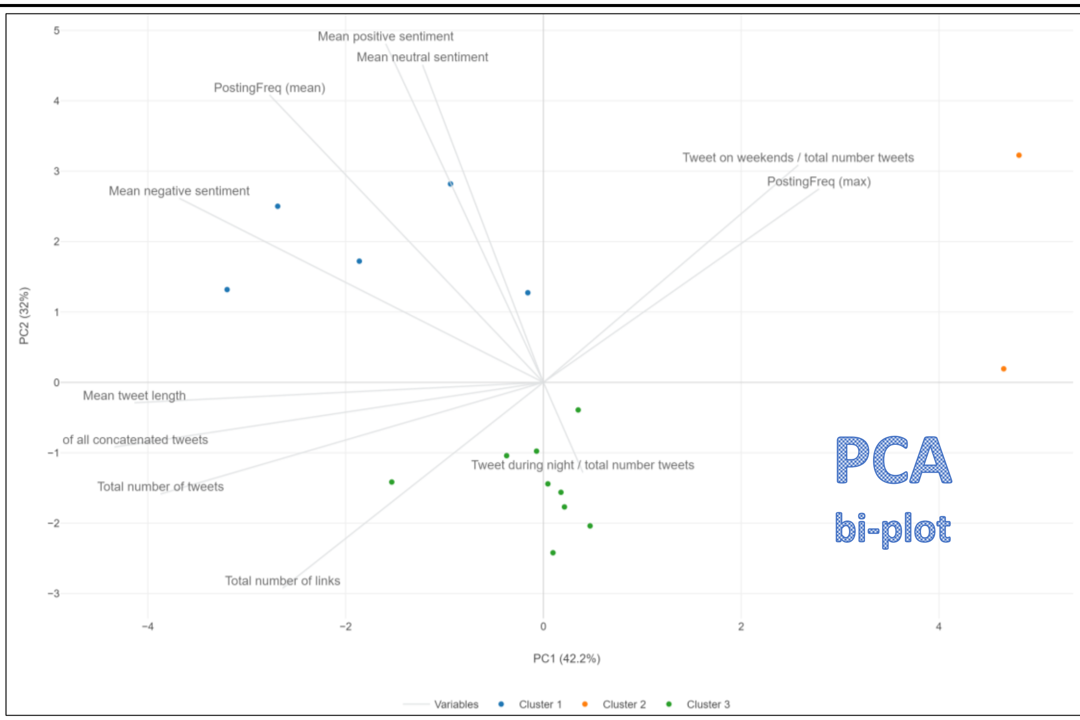
- Length of all concatenated tweets
- Mean negative sentiment
- Mean neutral sentiment
- Mean positive sentiment
- Mean tweet length
- PostingFreq (max)
- PostingFreq (mean)
- Total number of links
- Total number of tweets
- Tweet during night / total number tweets
- Tweet on weekends / total number tweets

Álvaro Figueira • VD • 2023 • 1ª ed.

40

We created 3 clusters

41



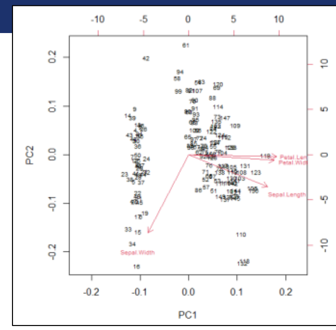
42

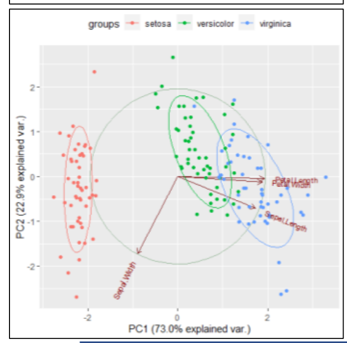
Creating bi-plots for PCA

```
# Load the iris dataset
data(iris)
# Exclude the Species (factor variable) column for PCA
iris.pca <- prcomp(iris[, -5], center = TRUE, scale. = TRUE)

# Create a bi-plot
biplot(iris.pca, cex = 0.6)

# Or we can use ggbiplot for a ggplot2-based biplot
# First we need to install it via devtools as it's not on CRAN
library(devtools)
devtools::install_github("vqv/ggbiplot")
# Now, load the library
library(ggbiplot)
# finally, the ggbiplot version
ggbiplot(iris.pca, obs.scale = 1, var.scale = 1,
          groups = iris$Species,  # grouping variable
          ellipse = TRUE,          # confidence area
          circle = TRUE) +         # correlation area
        theme(legend.direction = 'horizontal',
        legend.position = 'top')
```





Álvaro Figueira • VD • 2023 • 1ª ed.

43

Topic 7

Mapping values

Álvaro Figueira • VD • 2023 • 1ª ed.

44

Mapping Nominal Values to Numbers

- In case of ranked nominal values (e.g., air quality: bad, medium, good), there is a straightforward mapping: map each category into a consecutive integer
- In case of categorical values (e.g., car type), they can be transformed (expanded) into binary values, one column for each different category
  - This process is known as one-hot-encoding.

Note: One-hot encoding can significantly **increase the dimensionality** of the dataset if the categorical variable has many unique values. This can potentially lead to the "curse of dimensionality".
In such cases, it is wise to consider other encoding methods or dimensionality reduction.

Mapping Nominal Values to Numbers

Vehicle Name	Small/Sporty/ Compact/Large Sedan	Sports Car	SUV	Wagon	Minivan	Pickup	AWD	RWD	Retail Price	Dealer Cost	Engine Size (l)	Cyl	HP	City MPG	Hwy MPG	Weight	Wheel Base	Len	Width
Toyota 4Runner SR5 V6	0	0	1	0	0	0	0	0	27710	24801	4	6	245	18	21	4035	110	189	74
Toyota Avalon XL 4dr	1	0	0	0	0	0	0	0	26560	23693	3	6	210	21	29	3417	107	192	72
Toyota Avalon XLS 4dr	1	0	0	0	0	0	0	0	30920	27271	3	6	210	21	29	3439	107	192	72
Toyota Camry LE 4dr	1	0	0	0	0	0	0	0	19560	17558	2.4	4	157	24	33	3086	107	189	71
Toyota Camry LE V6 4dr	1	0	0	0	0	0	0	0	22775	20325	3	6	210	21	29	3296	107	189	71
Toyota Camry Solara SE 2dr	1	0	0	0	0	0	0	0	19635	17722	2.4	4	157	24	33	3175	107	193	72
Toyota Camry Solara SE V6 2dr	1	0	0	0	0	0	0	0	21965	19819	3.3	6	225	20	29	3417	107	193	72
Toyota Camry Solara SLE V6 2dr	1	0	0	0	0	0	0	0	26510	23908	3.3	6	225	20	29	3439	107	193	72
Toyota Camry XLE V6 4dr	1	0	0	0	0	0	0	0	25920	23125	3	6	210	21	29	3362	107	189	71
Toyota Celica GT-S 2dr	0	1	0	0	0	0	0	0	22570	20363									
Toyota Corolla CE 4dr	1	0	0	0	0	0	0	0	14085	13065									
Toyota Corolla LE 4dr	1	0	0	0	0	0	0	0	15295	13889									
Toyota Corolla S 4dr	1	0	0	0	0	0	0	0	15030	13650									
Toyota Echo 2dr auto	1	0	0	0	0	0	0	0	11560	10896									
Toyota Echo 2dr manual	1	0	0	0	0	0	0	0	10760	10144									
Toyota Echo 4dr	1	0	0	0	0	0	0	0	11290	10642									
Toyota Highlander V6	0	0	1	0	0	0	1	0	27930	24915	3.3	6	230	18	24	3935	107	185	72
Toyota Land Cruiser	0	0	1	0	0	0	1	0	54765	47986	4.7	8	325	13	17	5390	112	193	76
Toyota Matrix XR	0	0	0	1	0	0	0	0	16695	15156	1.8	4	130	29	36	2679	102	171	70
Toyota MR2 Spyder convertible 2dr	0	1	0	0	0	0	0	1	25130	22787	1.8	4	138	26	32	2195	97	153	67
Toyota Prius 4dr (gas/electric)	1	0	0	0	0	0	0	0	20510	18926	1.5	4	110	59	51	2890	106	175	68
Toyota RAV4	0	0	1	0	0	0	1	0	20290	18553	2.4	4	161	22	27	3119	98	167	68
Toyota Sequoia SR5	0	0	1	0	0	0	1	0	35695	31827	4.7	8	240	14	17	5270	118	204	78
Toyota Sienna CE	0	0	0	0	1	0	0	0	23495	21198	3.3	6	230	19	27	4120	119	200	77
Toyota Sienna XLE Limited	0	0	0	0	1	0	0	0	28800	25690	3.3	6	230	19	27	4165	119	200	77
Toyota Tacoma	0	0	0	0	0	1	0	1	12800	11879	2.4	4	142	22	27	2750	103	*	*
Toyota Tundra Access Cab V6 SR5	0	0	0	0	0	1	1	0	25935	23520	3.4	6	190	14	17	4435	128	*	*
Toyota Tundra Regular Cab V6	0	0	0	0	0	1	0	1	16495	14978	3.4	6	190	16	20	3925	128	*	*

Example of
one-hot-encoding

Mapping Nominal Values to Numbers

- For non-ranked values the problem is more complex (e.g., a person's name)
- If there is only one arbitrary nominal variable, we can use correspondence analysis
  - *Tuning procedure*: a **numerical value** can be assigned **using the other variables** to calculate a distance matrix, applying **MDS** (multidimensional scaling) to **calculate unidimensional coordinates**.

```
# Create the data frame
df <- data.frame(
        Animal = c("Lion", "Elephant", "Giraffe", "Bear", "Kangaroo", "Tiger"),
        Weight = c(190, 6000, 800, 500, 90, 220), # weights in kg
        Length = c(2.1, 3.3, 5.5, 1.8, 1.6, 2.3), # lengths in meters
        Diet = as.factor(c("Carnivore", "Herbivore", "Herbivore", "Omnivore",
                           "Herbivore", "Carnivore")))
# Perform one-hot encoding on the Diet variabl
df <- dummy_cols(df, select_columns = "Diet")
# Remove the original 'Diet' column
df$Diet <- NULL
# Compute the Euclidean distance matrix
dist_matrix <- dist(df[,-1]) # Exclude the 'Animal' column
# Perform MDS for one dimension (k=1)
mds_result <- cmdscale(dist_matrix, k = 1)
# Add the MDS result back to the original data frame
df$MDS_Coordinate <- mds_result
# Print the updated data frame
df
```

	Animal	Weight	Length	Diet	DietCarnivore	DietHerbivore	DietOmnivore	MDS_Coordinate
1	Lion	190	2.1	Carnivore	1	0	0	1110.0002
2	Elephant	6000	3.3	Herbivore	0	1	0	-4700.0001
3	Giraffe	800	5.5	Herbivore	0	1	0	499.9994
4	Bear	500	1.8	Omnivore	0	0	1	800.0002
5	Kangaroo	90	1.6	Herbivore	0	1	0	1210.0001
6	Tiger	220	2.3	Carnivore	1	0	0	1080.0002

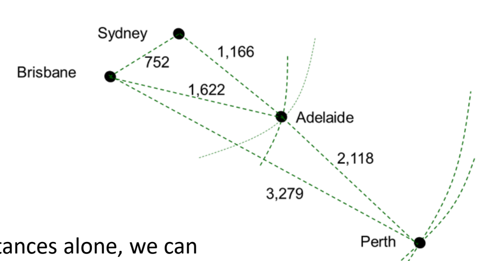
Multidimensional Scaling

- Multidimensional scaling (MDS) is a technique for visualizing distances between objects, where the distance is known between pairs of the objects.

The distance matrix below shows the distance, in kilometers, between four Australian cities.

	Sydney	Adelaide	Brisbane
Adelaide	1,166		
Brisbane	752	1,622	
Perth	3,279	2,118	3,606

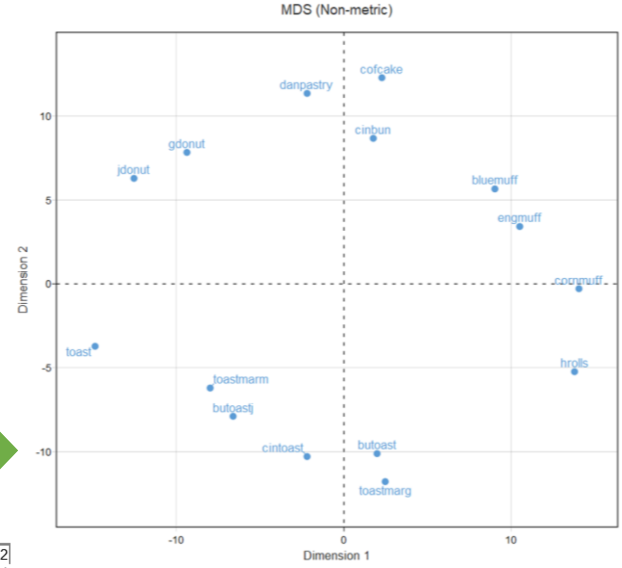
From these distances alone, we can reconstruct the map on the right.



Multidimensional Scaling (II)

- The distance matrix below shows the *perceived dissimilarities* between 15 breakfast baked goods, where a high number means that the subject rated them as being very dissimilar, and a lower number indicates the pair of baked breakfast goods are highly similar.

	toast	butoast	engmuff	jdonut	cintoast	bluemuff	hrolls	toastmarm	butoastj	toastmarg	cinbun	danpastry	gdonut	cofcake
butoast	15													
engmuff	25	15												
jdonut	3	24	22											
cintoast	14	3	17	22										
bluemuff	24	17	2	21	19									
hrolls	28	8	4	27	18	8								
toastmarm	7	7	20	11	6	18	23							
butoastj	8	6	21	12	5	19	22	2						
toastmarg	16	2	16	25	4	18	9	8	7					
cinbun	26	17	10	17	12	7	18	20	19	18				
danpastry	21	25	11	5	19	10	22	17	16	26	2			
gdonut	20	18	24	2	23	22	25	11	12	17	4	11		
cofcake	16	22	11	13	21	7	21	21	20	23	6	7	11	
cornmuff	27	11	3	26	16	4	5	25	24	12	12	16	24	16



Paul E. and Vithala R. Rao (1972), Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms. New York: Holt, Rinehart and Winston.

Multidimensional Scaling (III)

Note:

- When reading an MDS map, we can **consider only distances**. Unlike a geographic map, there is no concept of up or down, or north and south.
- All examples represent the same situation.

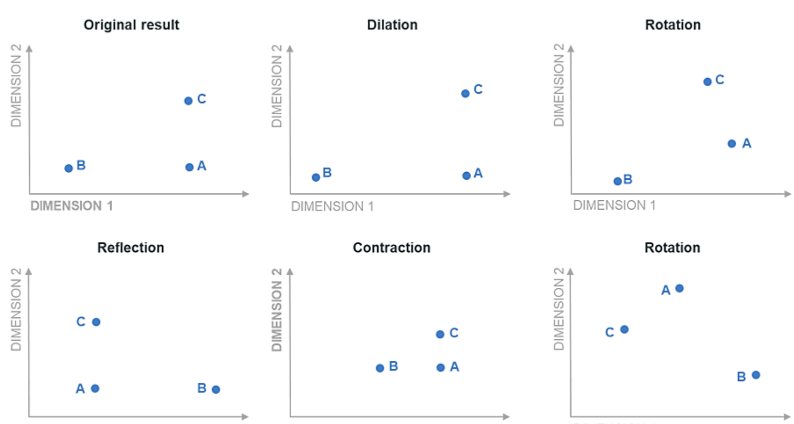


Figure from: Lehman, Donald (1989): Market Research and Analysis, 3rd Edition, Irwin.

Topic 8

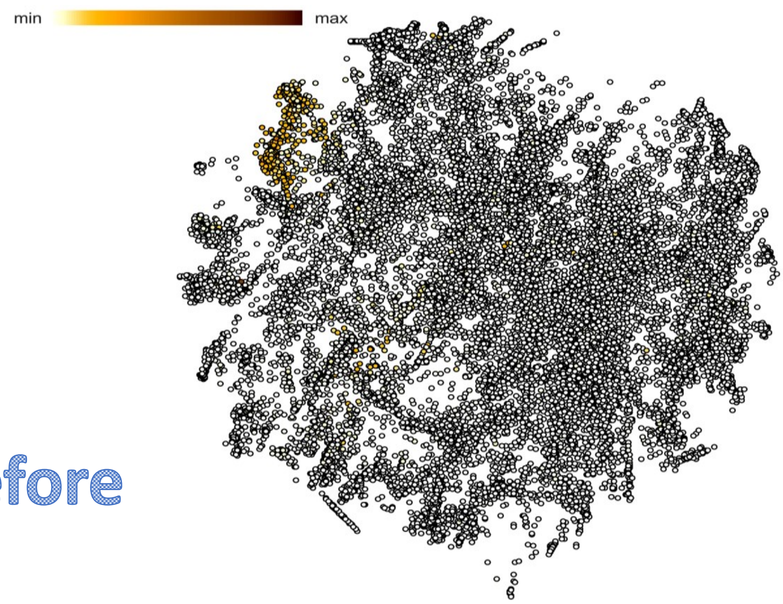
Aggregation and summarization

Aggregation and Summarization

- It can be useful to group data instances, using representatives for the groups (aggregation)
  - The average can be shown, or some other extra information, such as, the number of instances in a group (count)
  - Other aggregations are: std_dev, max, min, variance, etc.

- The core idea of **aggregation** is to provide information to help users to decide if a group needs to be further inspected:
  - To search for variability analysis, outlier detection, and others.

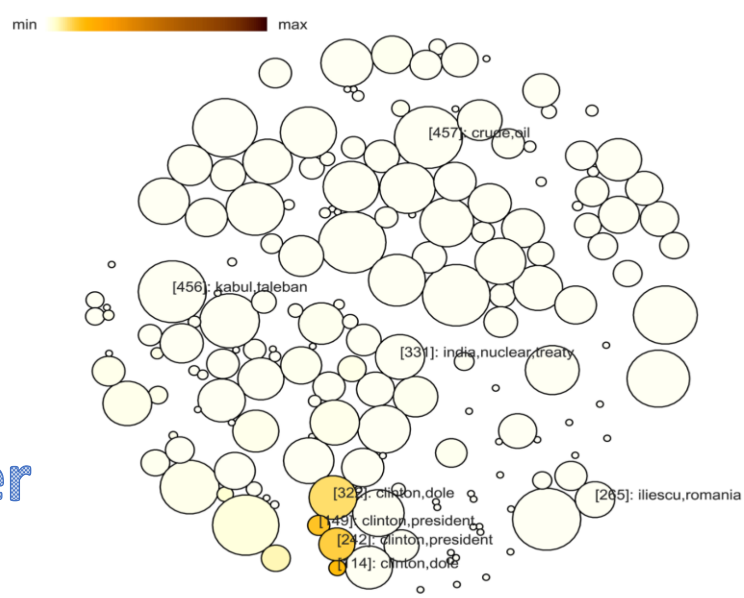
Aggregation and Summarization



before

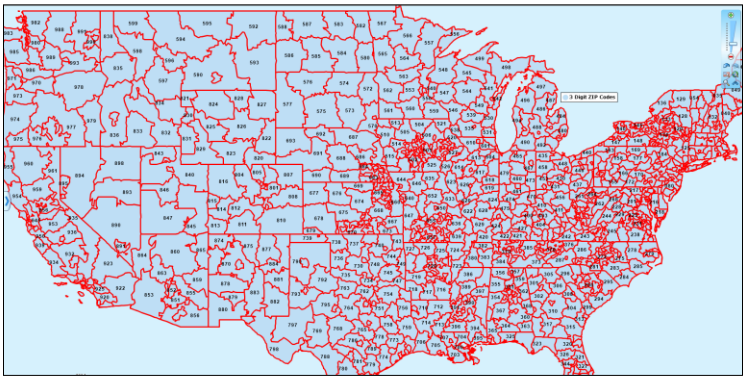
53

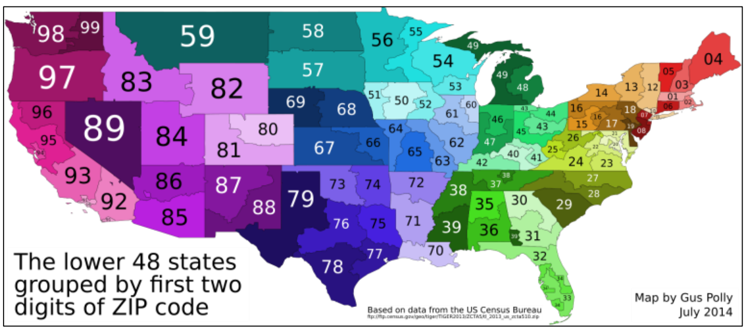
Aggregation and Summarization



after

54



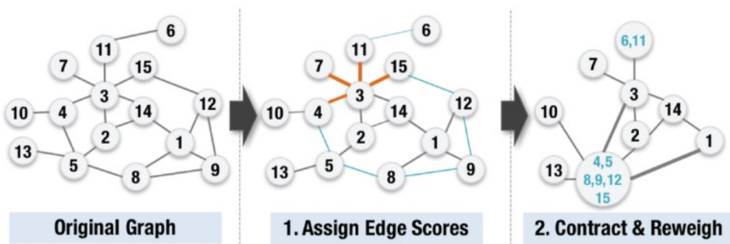


The ZIP codes in the US

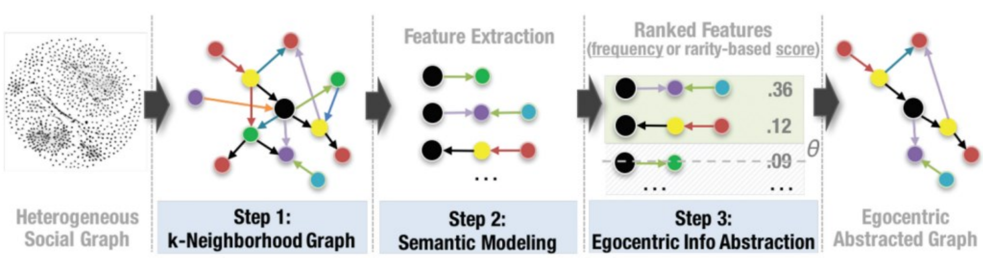
An aggregation view

55

Graph Agregation and Summarization



Graph aggregation

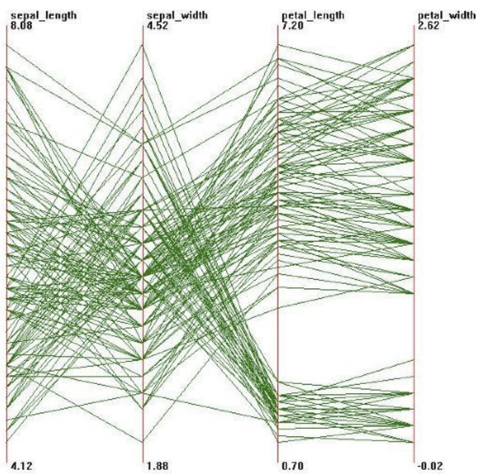


Graph summarization

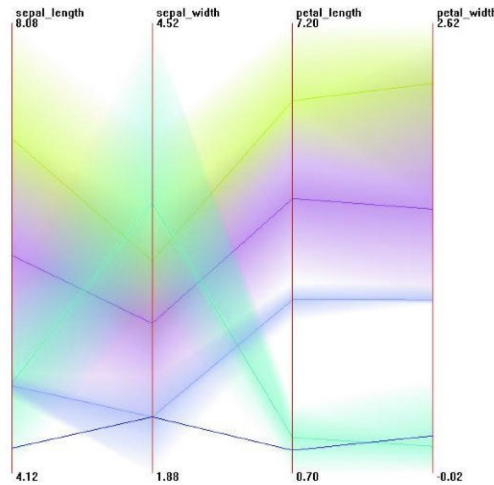
Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. 2018. Graph Summarization Methods and Applications: A Survey. ACM Comput. Surv. 51, 3, Article 62 (May 2019), 34 pages. https://doi.org/10.1145/3186727

56

Aggregation and Summarization



Original

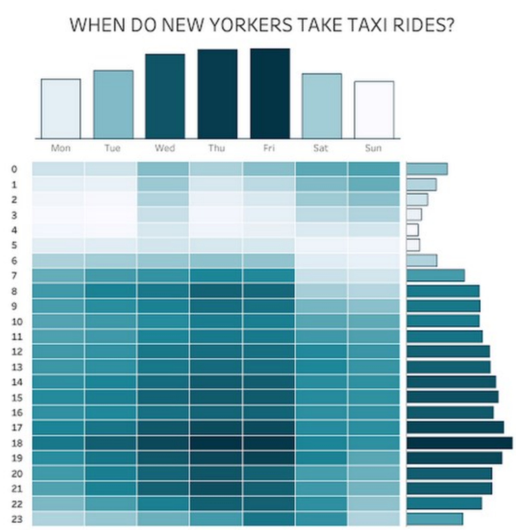


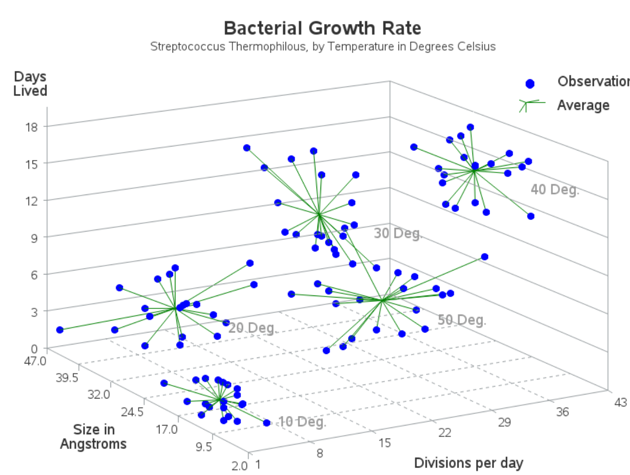
After aggregation

Álvaro Figueira • VD • 2023 • 1ª ed.

57

Aggregation





Álvaro Figueira • VD • 2023 • 1ª ed.

58

Final Observation

If the data were transformed through some process, this needs to be informed to the user or analyst!