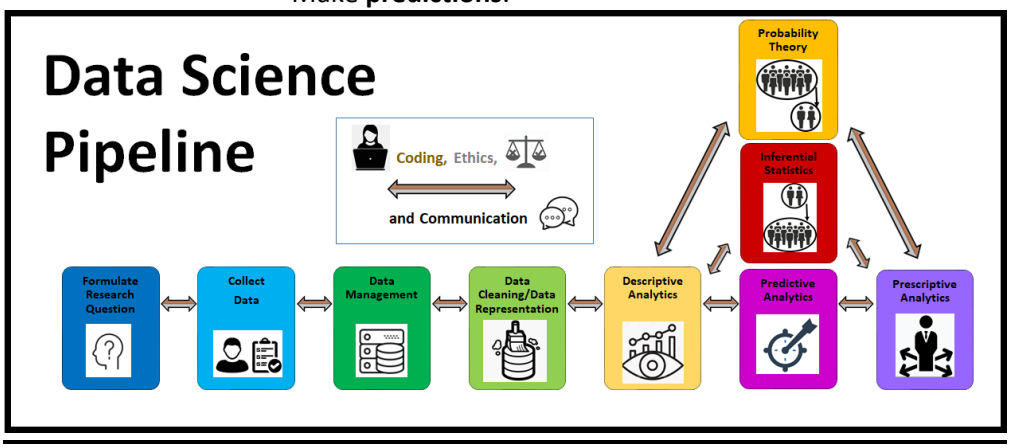
Unit 1: Introduction to Clustering

***Case Study*: Tripadvisor.com Ratings Data**

*Can you we tripadvisor.com review data to create **market segments** of customers in order to make **intelligent advertising campaign decisions?***

Purpose of this Lecture:

- Introduce a common clustering algorithm to provide an example of how a clustering algorithm might **fit into the "full data science pipeline."**
- Introduce a common clustering algorithm to introduce the **types of questions** we might ask when making the following analysis choices.
    - Is a clustering algorithm **useful** to apply to the given dataset?
    - **Which clustering algorithm** should we use for a given dataset?
    - **Which parameters** for the selected clustering algorithm should be chosen?
    - Did the clustering algorithm discover **any meaningful clusters**?
    - How can we use the results of this clustering algorithm to:
        - Perform **data cleaning**
        - Discover **actionable insights** about the data
        - Make **predictions**.



Summary of Concepts:

- Why do we need clustering?
- K-means clustering algorithm
    - Input:
        - Parameter: k
        - Numerical data
    - Output:
        - Partition of objects
    - Algorithm properties:
        - *Inertia*: Objective function of the k-means clustering algorithm
        - Non-deterministic nature of the k-means clustering algorithm
        - Local optimality of the k-means clustering algorithm
        - Types of datasets that work well with k-means clustering algorithm.
    - Choosing the right parameters: Elbow-method for choosing the right k.

- Benefits/Drawbacks of k-means clustering algorithms

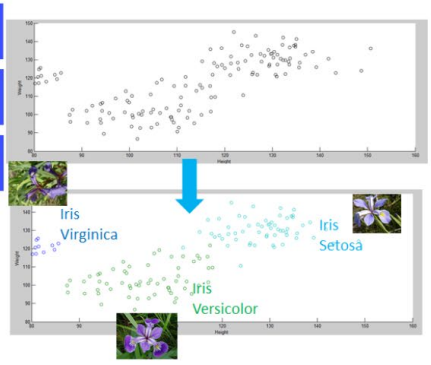
Introduction to Clustering

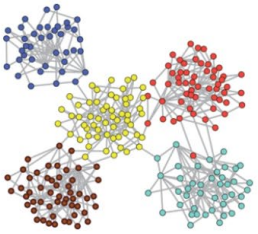
with the k-Means Clustering Algorithm

Motivation | Collect Data (sample) | Descriptive Analytics (sample) | Predictive Analytics (new data) | Prescriptive Analytics (decisions)

Clustering

- Understanding (2d data)
- Realistic Data (>2d data)
- k-Means Clustering
  - Running the Algorithm
  - Output
  - Assessing Performance
  - Benefits/Drawbacks
  - When to use k-means
  - Assessing Data Assumptions





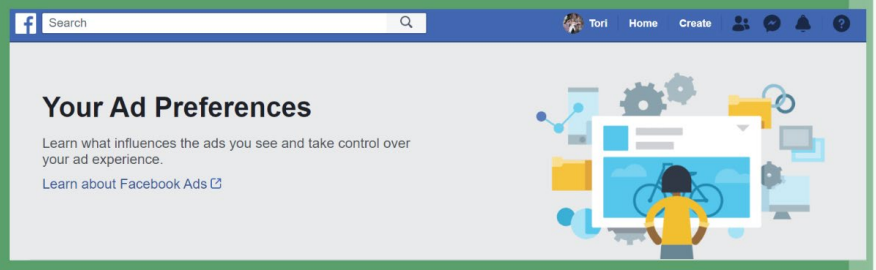
Motivation

vs.

Motivation

Market Segmentation:

Market segmentation is the process of dividing a market of potential customers into groups, or segments, based on different characteristics. The segments created are composed of consumers who will respond similarly to marketing strategies and who share traits such as similar interests, needs, or locations.



The Chronicle of Higher Education

Game of Thrones

Well-being

Data science

Baby boomers

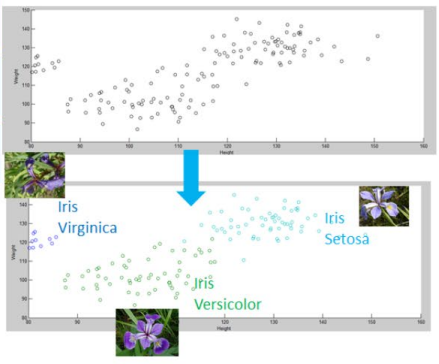
Introduction to Clustering

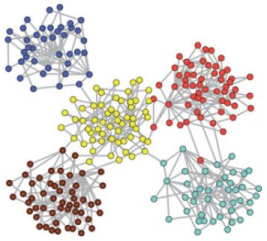
with the k-Means Clustering Algorithm

Motivation	Collect Data (sample)	Descriptive Analytics (sample)	Predictive Analytics (new data)	Prescriptive Analytics (decisions)

Clustering

- Understanding (2d data)
- Realistic Data (>2d data)
- k-Means Clustering
  - Running the Algorithm
  - Output
  - Assessing Performance
  - Benefits/Drawbacks
  - When to use k-means
  - Assessing Data Assumptions





Goal: How best to **group** a set of objects:

- The **similarity** of the objects in the **same cluste**r is **maximized** while
- The **similarity** of objects in **different clusters** is **minimized**

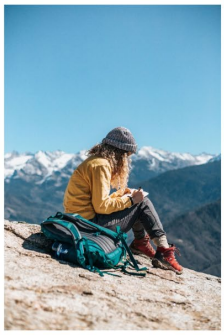
Collect a Sample

Source: Tripadvisor.com

Sample Size: n = 980

Data: This data set is populated by crawling TripAdvisor.com. Reviews on destinations in 10 categories mentioned across East Asia are considered. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0) and average rating is used against each category per user.

https://archive.ics.uci.edu/ml/datasets/Travel+Reviews



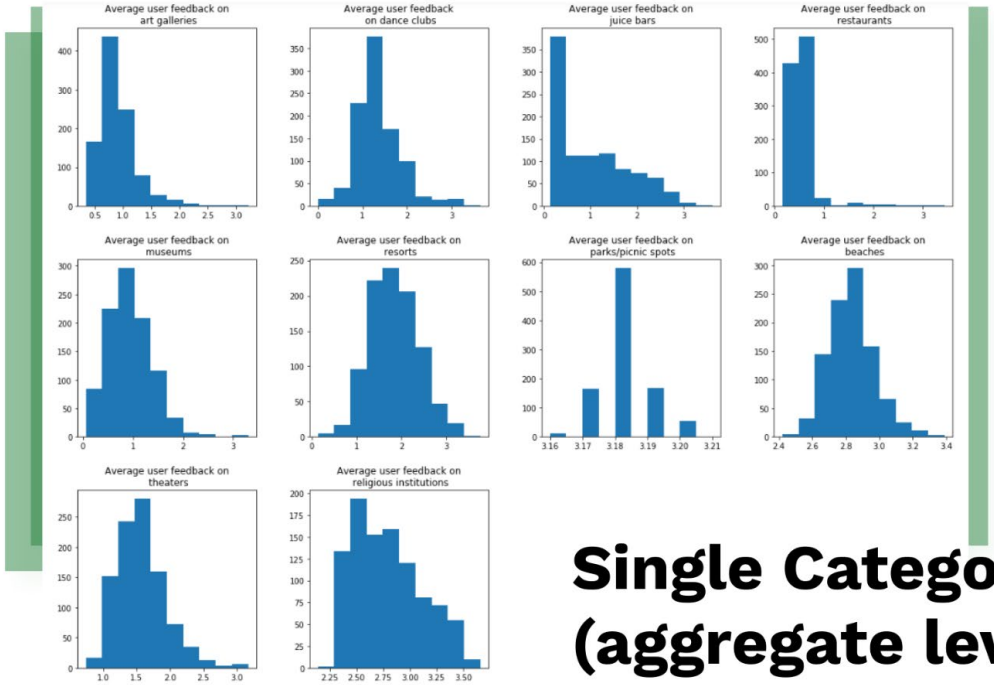
10 features:

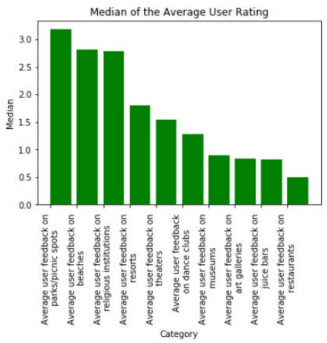
Average user feedback on art galleries
Average user feedback on dance clubs
Average user feedback on juice bars
Average user feedback on restaurants
Average user feedback on museums
Average user feedback on resorts
Average user feedback on parks/picnic spots
Average user feedback on beaches
Average user feedback on theaters
Average user feedback on religious institutions

10 features:

Average user feedback on art galleries
Average user feedback on dance clubs
Average user feedback on juice bars
Average user feedback on restaurants
Average user feedback on museums
Average user feedback on resorts
Average user feedback on parks/picnic spots
Average user feedback on beaches
Average user feedback on theaters
Average user feedback on religious institutions

Exploratory Data Analysis on the Sample

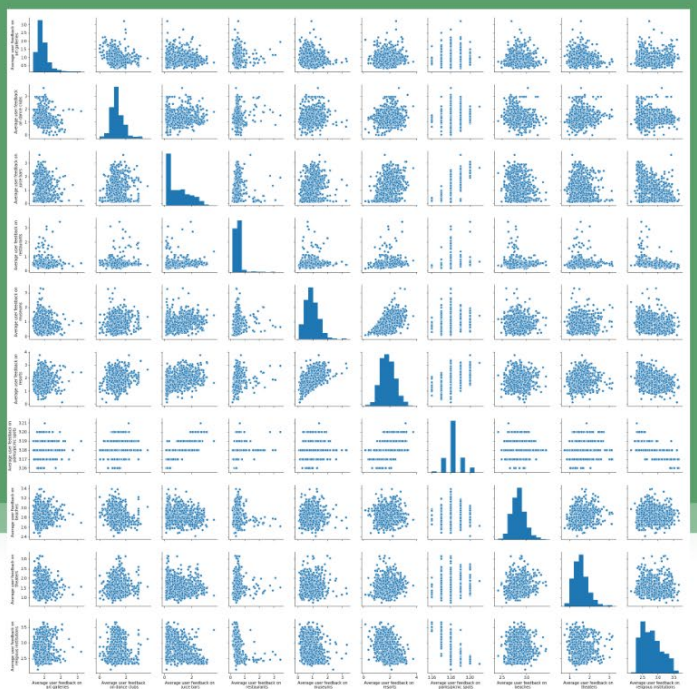




Single Category Analysis (aggregate level)

Exploratory Data Analysis on the Sample

Relationship Between Two Categories Analysis (aggregate level)



Exploratory Data Analysis on the Sample

Relationship Between Two Categories Analysis (aggregate level)

Does there exist **clusters** of individuals with high similarity?

What makes individuals in these clusters most similar? How do we use this to advertise to them?

Understanding Clustering (an ideal case 2D)

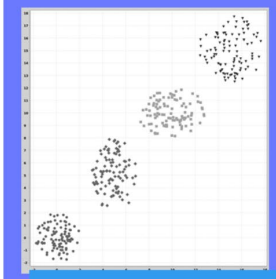
Original Data

Each object has *only 2* attributes

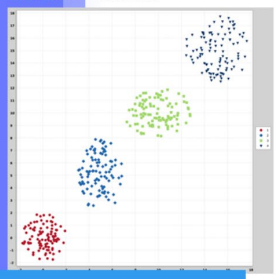
1000 Objects

	Attribute 1	Attribute 2
Tissue Sample (Object) 1:	7.59	1.69
Tissue Sample (Object) 2:	10.62	0.88
Tissue Sample (Object) 3	8.22	1.96
Tissue Sample (Object) 4	10.68	1.46
Tissue Sample (Object) 5	14.55	0.53
Tissue Sample (Object) 6	7.53	0.18
Tissue Sample (Object) 7	7.28	2.93
Tissue Sample (Object) 8	9.43	1.05
Tissue Sample (Object) 9	6.44	2.56
Tissue Sample (Object) 10	27.02	0.8
Tissue Sample (Object) 11	9.16	0.15
Tissue Sample (Object) 12	6.77	1.62
Tissue Sample (Object) 13	12.51	7.03
Tissue Sample (Object) 14	1.38	2.03
Tissue Sample (Object) 15	15.57	0.5
Tissue Sample (Object) 16	8.14	2.45
Tissue Sample (Object) 17	6.67	0.17
Tissue Sample (Object) 18	6.17	1.5
Tissue Sample (Object) 19	11.87	0.09
Tissue Sample (Object) 20	16.56	0.71
Tissue Sample (Object) 21	11.29	0.92
Tissue Sample (Object) 22	10.2	-0.24
Tissue Sample (Object) 23	10.99	1.32
Tissue Sample (Object) 24	11.64	0.39
Tissue Sample (Object) 25	6.62	1.02
Tissue Sample (Object) 26	9.18	-0.27
Tissue Sample (Object) 27	7.56	0.19
Tissue Sample (Object) 28	18.02	0.23
Tissue Sample (Object) 29	10.24	-0.31
...		
Tissue Sample (Object) 993	19.74	0.05
Tissue Sample (Object) 994	10.84	0.48
Tissue Sample (Object) 995	10.9	0.47
Tissue Sample (Object) 996	9.08	0
Tissue Sample (Object) 997	24.73	-0.51
Tissue Sample (Object) 998	16.34	0.43
Tissue Sample (Object) 999	15.9	1.31
Tissue Sample (Object) 1000	13.22	-0.3

Step 1: Graph the data



Step 2: Visually inspect to assign cluster labels



We can visually see:

- the data is **"clusterable"** (there exist natural groupings)
- there are four **"natural" groups**
- which **object belongs** in which group.

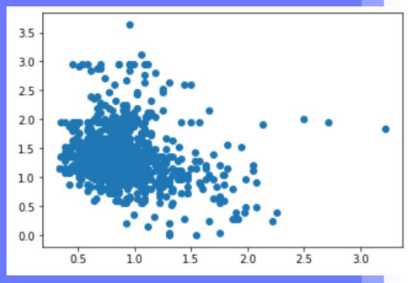
Understanding Clustering (an ideal case 2D)

Original Data

Each object has *only 2* attributes

1000 Objects

	Attribute 1	Attribute 2
Tissue Sample (Object) 1:	7.59	1.69
Tissue Sample (Object) 2:	10.62	0.88
Tissue Sample (Object) 3	8.22	1.96
Tissue Sample (Object) 4	10.68	1.46
Tissue Sample (Object) 5	14.55	0.53
Tissue Sample (Object) 6	7.53	0.18
Tissue Sample (Object) 7	7.28	2.93
Tissue Sample (Object) 8	9.43	1.05
Tissue Sample (Object) 9	6.44	2.56
Tissue Sample (Object) 10	27.02	0.8
Tissue Sample (Object) 11	9.16	0.15
Tissue Sample (Object) 12	6.77	1.62
Tissue Sample (Object) 13	12.51	7.03
Tissue Sample (Object) 14	1.38	2.03
Tissue Sample (Object) 15	15.57	0.5
Tissue Sample (Object) 16	8.14	2.45
Tissue Sample (Object) 17	6.67	0.17
Tissue Sample (Object) 18	6.17	1.5
Tissue Sample (Object) 19	11.87	0.09
Tissue Sample (Object) 20	16.56	0.71
Tissue Sample (Object) 21	11.29	0.92
Tissue Sample (Object) 22	10.2	-0.24
Tissue Sample (Object) 23	10.99	1.32
Tissue Sample (Object) 24	11.64	0.39
Tissue Sample (Object) 25	6.62	1.02
Tissue Sample (Object) 26	9.18	-0.27
Tissue Sample (Object) 27	7.56	0.19
Tissue Sample (Object) 28	18.02	0.23
Tissue Sample (Object) 29	10.24	-0.31
...		
Tissue Sample (Object) 993	19.74	0.05
Tissue Sample (Object) 994	10.84	0.48
Tissue Sample (Object) 995	10.9	0.47
Tissue Sample (Object) 996	9.08	0
Tissue Sample (Object) 997	24.73	-0.51
Tissue Sample (Object) 998	16.34	0.43
Tissue Sample (Object) 999	15.9	1.31
Tissue Sample (Object) 1000	13.22	-0.3



We can visually see:

- the data is NOT **"clusterable"** (there do not exist natural groupings in just this data)
- all **objects belong** in the same group.

Realistic Data (>2D data)

Original Data

10 attributes (dimensions)

980 objects

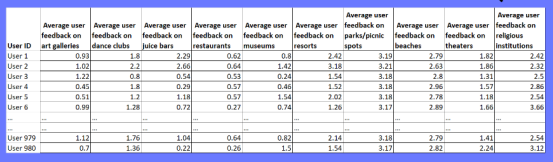
User ID	Average user feedback on art galleries	Average user feedback on dance clubs	Average user feedback on juice bars	Average user feedback on restaurants	Average user feedback on museums	Average user feedback on resorts	Average user feedback on parks/picnic spots	Average user feedback on beaches	Average user feedback on theaters	Average user feedback on religious institutions
User 1	0.93	1.8	2.29	0.62	0.8	2.42	3.19	2.79	1.82	2.42
User 2	1.02	2.2	2.66	0.64	1.42	3.18	3.21	2.63	1.86	2.32
User 3	1.22	0.8	0.54	0.53	0.24	1.54	3.18	2.8	1.31	2.5
User 4	0.45	1.8	0.29	0.57	0.46	1.52	3.18	2.96	1.57	2.86
User 5	0.51	1.2	1.18	0.57	1.54	2.02	3.18	2.78	1.18	2.54
User 6	0.99	1.28	0.72	0.27	0.74	1.26	3.17	2.89	1.66	3.66
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
User 979	1.12	1.76	1.04	0.64	0.82	2.14	3.18	2.79	1.41	2.54
User 980	0.7	1.36	0.22	0.26	1.5	1.54	3.17	2.82	2.24	3.12

We can NOT visually see:
- if the data is **"clusterable"** (there exist natural groupings)
- and **how many "natural" groupings** there would be.
- which **user belongs** in which grouping.

Use a **clustering algorithm**!

K-Means Clustering

Input:

1. Desired number of clusters (ex: k=3)
2. 

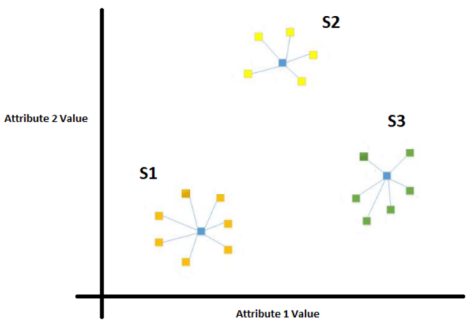
User ID	Average user feedback on art galleries	Average user feedback on dance clubs	Average user feedback on juice bars	Average user feedback on restaurants	Average user feedback on museums	Average user feedback on resorts	Average user feedback on parks/picnic spots	Average user feedback on beaches	Average user feedback on theaters	Average user feedback on religious institutions
User 1	0.93	1.8	2.29	0.62	0.8	2.42	3.19	2.79	1.82	2.42
User 2	1.02	2.2	2.66	0.64	1.42	3.18	3.21	2.63	1.86	2.32
User 3	1.22	0.8	0.54	0.53	0.24	1.54	3.18	2.8	1.31	2.5
User 4	0.45	1.8	0.29	0.57	0.46	1.52	3.18	2.96	1.57	2.86
User 5	0.51	1.2	1.18	0.57	1.54	2.02	3.18	2.78	1.18	2.54
User 6	0.99	1.28	0.72	0.27	0.74	1.26	3.17	2.89	1.66	3.66
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
User 979	1.12	1.76	1.04	0.64	0.82	2.14	3.18	2.79	1.41	2.54
User 980	0.7	1.36	0.22	0.26	1.5	1.54	3.17	2.82	2.24	3.12

Goal:

Informally: Find a grouping of the objects such that the sum of distances between the each object and the mean of the group that is assigned to (centroid) is minimized.

Technically:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$



Decision: S1, S2,...,Sk

How to find a solution to the k-means clustering problem?

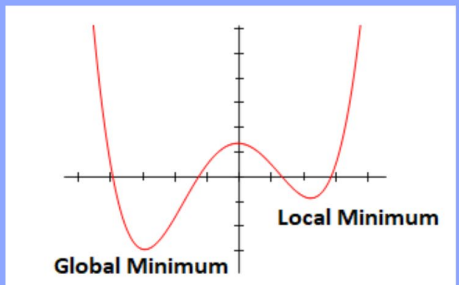
Best Case: Global minimum

Decision: S1, S2,...,Sk

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

➡ NP-Hard

Heuristic Algorithm: Find a local minimum.



How to find a solution to the k-means clustering problem?

Best Case: Global minimum

Decision: S1, S2,...,Sk

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

➡ NP-Hard

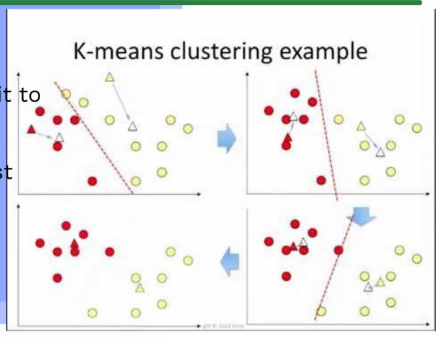
Heuristic Algorithm: Find a local minimum.

Step 1: **Randomly create k centroid** points.

Step 2: For each object in the data set, **assign** it to the centroid that is closest.

If all of these new (centroid, object) assignments are exactly the same as the last (centroid, object) assignment made by the previous iteration of the algorithm, STOP. Otherwise, proceed to step (3).

Step 3: **Find the mean of each cluster** created from step (2). These mean points become the new centroids points. Go back to step (2).

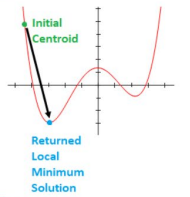


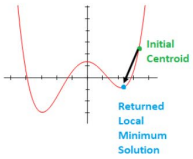
K-means Clustering Algorithm Output

User ID	Assigned Cluster Label
User 1	0
User 2	0
User 3	1
User 4	1
User 5	2
User 6	1
...	...
User 979	1
User 980	2

Warning: K-means algorithm:
- Begins with a random initial start point each time and
- It only returns a "local optimal solution" found from this random initial starting point.

Translation: K-means may return different clustering results, each time you run it.
- Some results may be better than others.
- Some results may represent a different "facet" of how the data could be grouped.



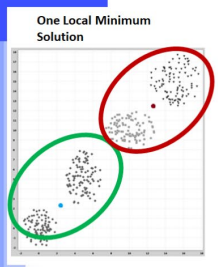


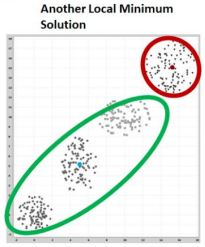
K-means Clustering Algorithm Output

User ID	Assigned Cluster Label
User 1	0
User 2	0
User 3	1
User 4	1
User 5	2
User 6	1
...	...
User 979	1
User 980	2

Warning: K-means algorithm:
- Begins with a random initial start point each time and
- It only returns a "local optimal solution" found from this random initial starting point.

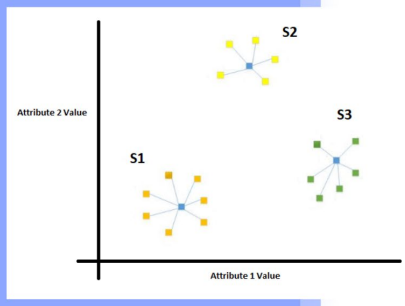
Translation: K-means may return different clustering results, each time you run it.
- Some results may be better than others.
- Some results may represent a different "facet" of how the data could be grouped.





Assessing Performance of Results

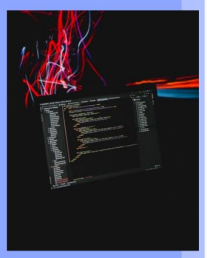
Returned Clusters (ex):
- S1 = {user 1, user 2,... }
- S2 = {user 3, user 4,...}
- S3 = {user 5, user 7,...}



Inertia of the Clustering

Benefits/Drawbacks of k-means clustering (over other clustering algorithms)

Benefits:
- Fast algorithm.
- Good especially if:
  - # of objects if really large and/or
  - # of attributes is really large.
- One of the easiest to understand.

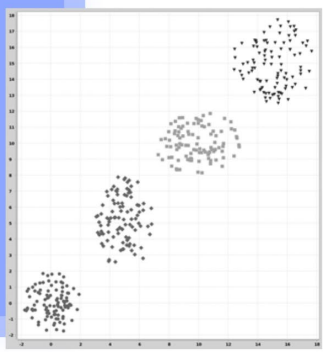


Drawbacks:
- Works well with only some types of data.
- Need to know a suitable number of clusters to request in advance (see elbow method in Python analysis)
- Different outputs (some better/some worse) may be returned each time it is run.

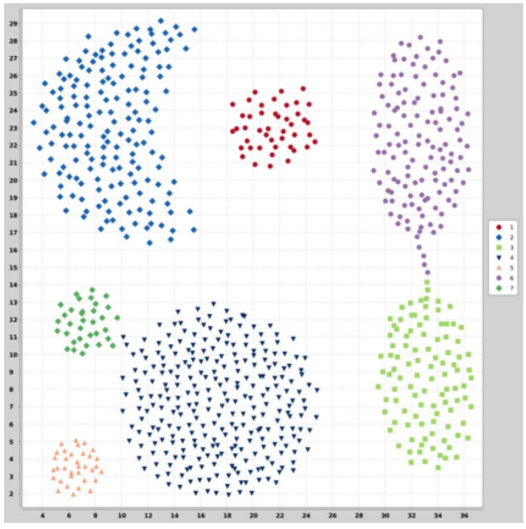
When to use the k-means algorithm

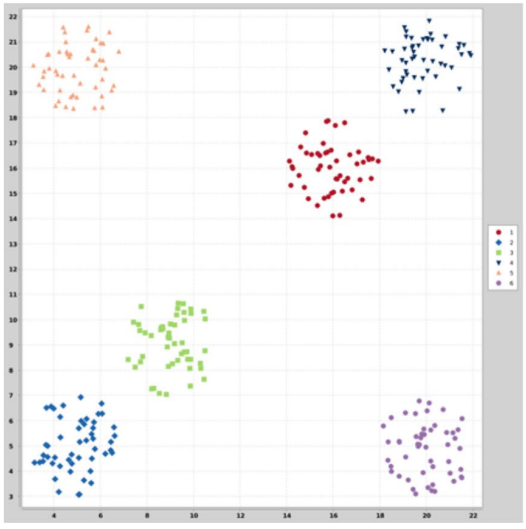
Assumptions about the data
The K-means algorithm works best for data when "the intended clustering" of the data has the following properties:

- Each cluster has roughly the same number of objects.
- The clusters are spherical.
- There is good separation between the clusters.
- You know the right number of clusters to ask for.
- Attributes are non-categorical.
- Data does not have a lot of noise or outliers.



Question: Which data set would we expect for the k-means clustering algorithm to have the best performance with?

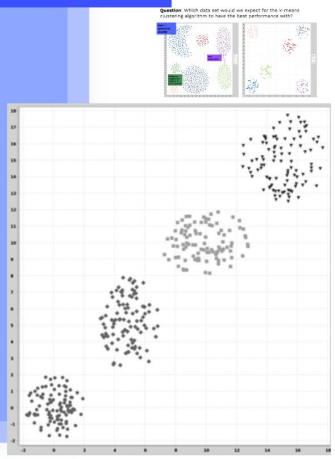




When to use the k-means algorithm

Assumptions about the data
The K-means algorithm works best for data when "the intended clustering" of the data has the following properties:

- Each cluster has roughly the same number of objects.
- The clusters are spherical.
- There is good separation between the clusters.
- You know the right number of clusters to ask for.
- Attributes are non-categorical.
- Data does not have a lot of noise or outliers.



Use "Elbow Method."

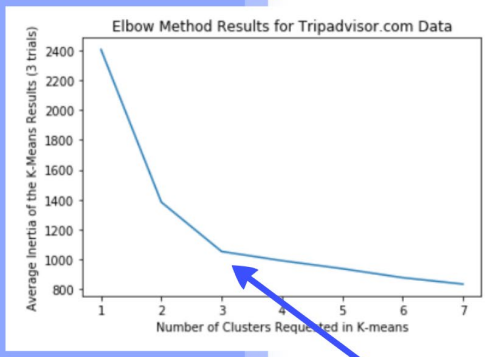
Assessing Input Data Assumptions (with the Elbow Method)

Elbow Method

- **FOR** k = 1 to K:
  - **Cluster the data** several times into **k clusters**.
  - Calculate the **average inertia** of these resulting clusterings.
  - **Plot** (k, average inertia).

Interpreting the Final Plot:

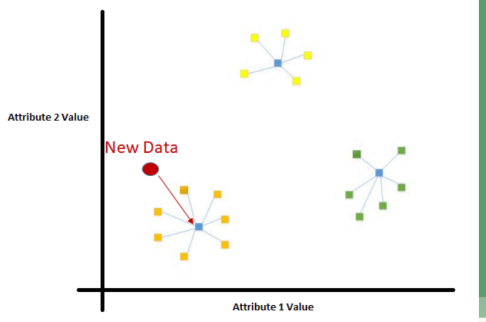
- More "dramatic" the "elbow" the more "clusterable" the data is.
- Ideal number of clusters = where the plot levels off.



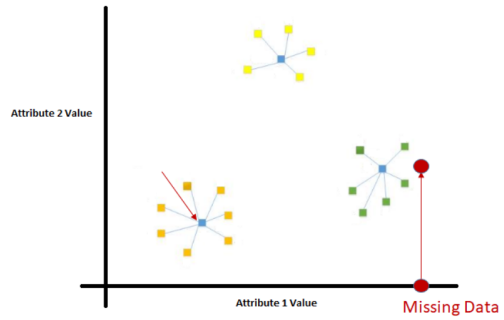
- Relatively clusterable data
- Ideal k=3

Predictions: Assigning New Data to Clusters or Predicting Missing Values

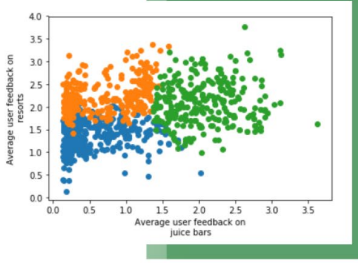
Assigning New Data to Clusters after Running Algorithm

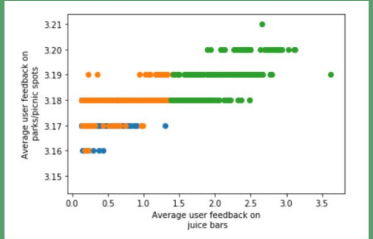


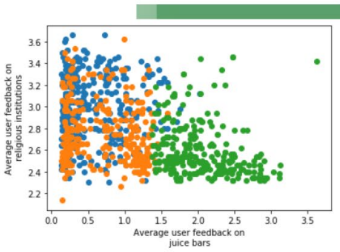
Predicting Missing Attribute Data (Collaborative Filtering)

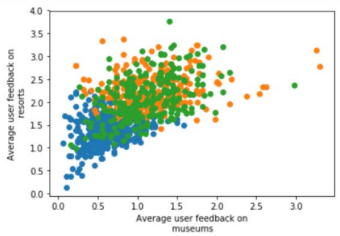


Using Clusters (Market Segments) to Make Informed Advertising Decisions

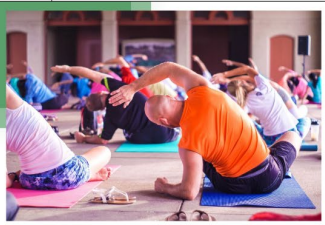








	Characteristics
Cluster 1	**Higher Socio-economic Status** Higher preference for resorts and museums. Lower preference of juice bars, parks, and picnic spots.
Cluster 2	**Lower Socio-economic Status** Lower preference for resorts and museums. Lower preference of juice bars, parks, and picnic spots.
Cluster 3	**Non-religious Health-Focused** Higher preference for juice bars, parks, and picnic spots. Lower preference for religious institutions



References

- https://github.com/deric/clustering-benchmark
- Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- https://en.wikipedia.org/wiki/K-means_clustering
- Müller-Linow M, Hilgetag CC, Hütt M-T (2008) Organization of Excitable Dynamics in Hierarchical Biological Networks. PLoS Comput Biol 4(9): e1000190. https://doi.org/10.1371/journal.pcbi.1000190
- https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
- https://www.youtube.com/watch?v=_aWzGGNrcic
- Renjith, Shini & Sreekumar, A. & Jathavedan, M.. (2018). Evaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism Domain. Proceedings of the IEEE. 2018 Recent Advances in Intelligent Computational Systems (RAICS). 127-131. 10.1109/RAICS.2018.8635080.
- https://trackmaven.com/marketing-dictionary/market-segmentation/