

Clustering

 ... reviewing machine learning literature about Clustering

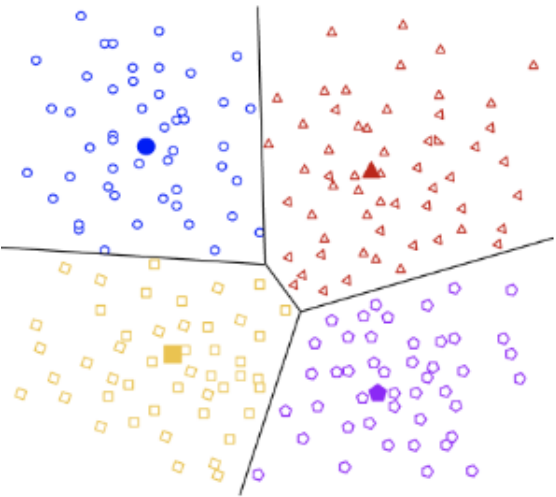
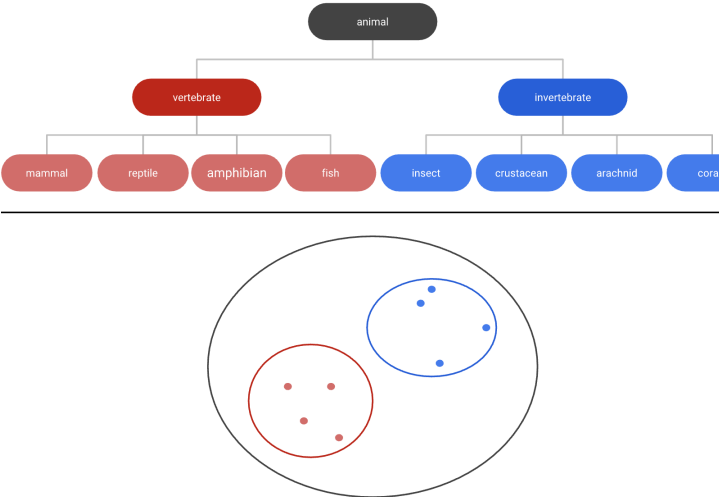
clustering

- unsupervised learning
- groups data points based on their similarities
- each group is called a cluster and contains data points with high similarity and low similarity with data points in other clusters
- Clustering is divided into two subgroups based on the assignment of data points to clusters:
 - Hard : Each data point is assigned to exactly one cluster. One example is k-means clustering.
 - Soft : Each data point is assigned a probability or likelihood of being in a cluster. One example is expectation-maximisation (EM) algorithm.

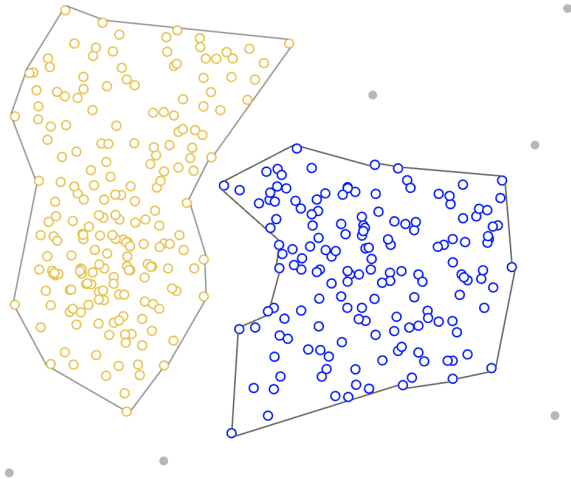
clustering vs classification

- Classification and clustering are two methods of **pattern identification used in machine learning**.
- Classification uses **predefined classes** in which objects are assigned, while clustering **identifies similarities between objects**, which it **groups** according to those characteristics in common and which differentiate them from other groups of objects. These groups are known as "**clusters**".
- classification : supervised learning, binary classification

Type of clustering algorithms

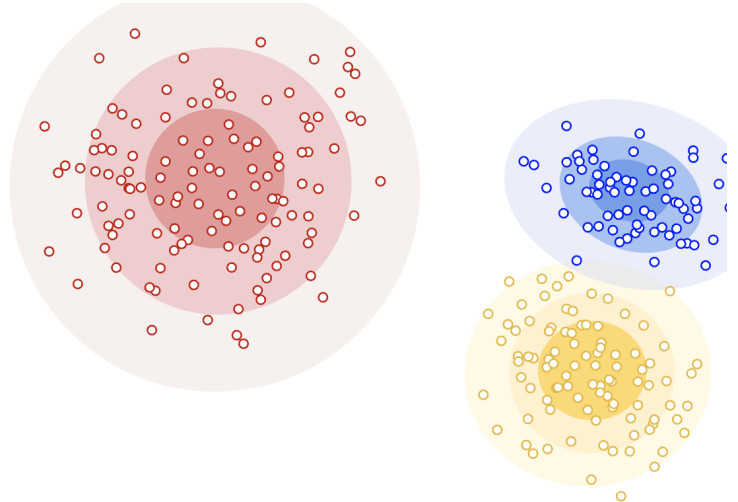
nature of data	
<div><h3>centroid-based clustering</h3><ul style="list-style-type: none">▪ It is an iterative clustering algorithm in which similarity is based on the proximity of a data point to the centroids of the clusters.▪ K-means clustering is one example of this model.▪ It needs a number of clusters before running and then divides data points into these many clusters iteratively.▪ Therefore, to use k-means, users should acquire some prior knowledge about the dataset.▪ Centroid-based algorithms are efficient but sensitive to initial conditions and outliers.</div>	<div><h3>connectivity-based clustering</h3><ul style="list-style-type: none">▪ This model assigns higher similarity to data points which are closer in one or more dimensional space than those points which are farther away.▪ There are two approaches - first, it categorises all data points into different clusters and then merges the data points in relation to the distances among them.▪ Second, it categorises all data points into one single cluster and then partitions them into different clusters as the distance increases.▪ This model is easy to understand but has problems in handling large datasets▪ One example is hierarchical clustering and its variants.</div>

density-based clustering



- This model searches one or multi-dimensional space for dense regions (having a large number of data points in a small region)
- A popular example of a density model is DBSCAN.
- Density-based clustering connects areas of high example density into clusters.
- This allows for arbitrary-shaped distributions as long as dense areas can be connected.
- These algorithms have difficulty with data of varying densities and high dimensions.
- Further, by design, these algorithms do not assign outliers to clusters.

distribution-based clustering



- This clustering approach assumes data is composed of distributions, such as Gaussian distributions.
- As distance from the distribution's centre increases, the probability that a point belongs to the distribution decreases.
- The bands show that decrease in probability.
- When you do not know the type of distribution in your data, you should use a different algorithm.

algorithm	description
-----------	-------------

k-means clustering

- data is grouped in terms of characteristics and similarities
- k represents the number of clusters
- key concept
 - squared Euclidean distance :

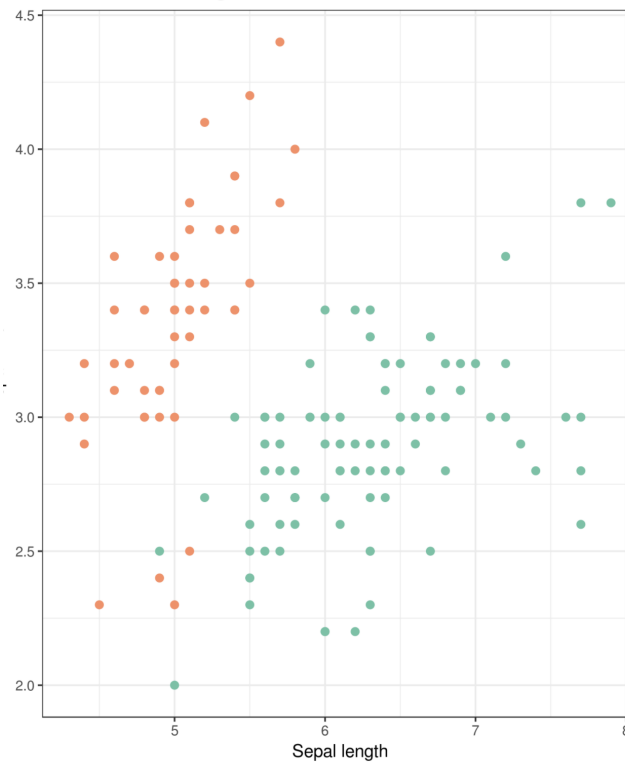
$$d(x, y)^2 = \sum_{j=1}^m (x_j - y_j)^2 = \|x - y\|_2^2$$

- cluster inertia (sum of squared errors) :

$$SSE = \sum_{i=1}^n \sum_{j=1}^k w^{(i,j)} \|x^{(i)} - \mu^{(j)}\|_2^2$$

- algorithm steps
 - choose the value of k (the number of desired clusters)
 - methods: field knowledge, business decision, elbow method, silhouette value
 - elbow method is the most commonly used
 - select centroids randomly
 - use the Euclidean distance (between centroids and data points) to assign every data point to the closest cluster
 - recalculate the centres of all clusters
 - repeat 3-4 until there is no further change
- advantages
 - efficient in terms of computation
 - can be implemented easily
- disadvantages
 - not effective for spherically distributed data.
 - the random selection of initial centroids may make some outputs to be different and this may affect the entire algorithm process

K-means Clustering in Iris Data

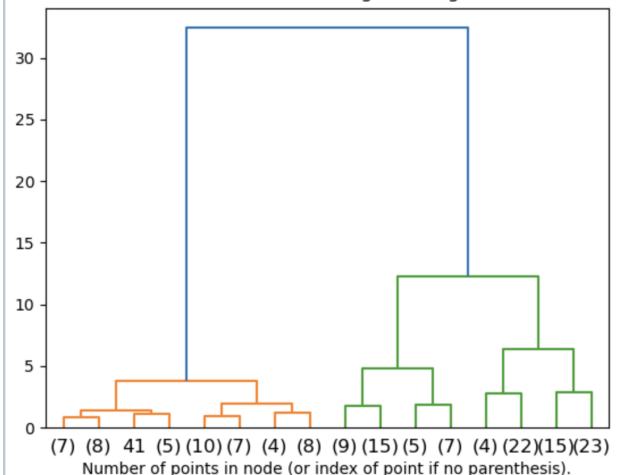


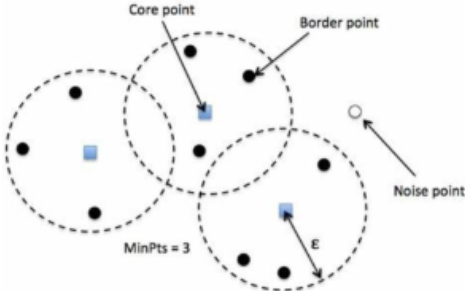
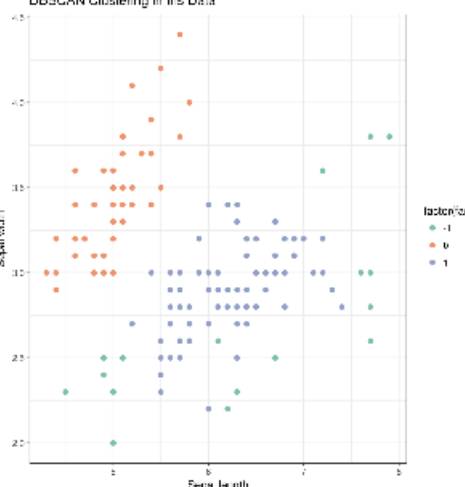

- silhouette : a method of interpretation and validation of consistency
- silhouette value : a measure of how similar an object is to its own cluster (separation).

hierarchical clustering

- is used when constructing a hierarchy of clusters
- starts by allocating each point of data to its cluster
- ends only if there is only one cluster left
- i.e. dendrogram
- algorithm steps
 - allocate each data point to its cluster
 - use Euclidean distance to locate two closest clusters, then merge these clusters to form one cluster
 - determine the distance between clusters that are near each other, then combine the nearest clusters until we have grouped all the data items to form a single cluster
- advantages
 - provide meaningful information
 - not require the number of clusters to be specified
 - resourceful for the construction of dendrograms
- disadvantages
 - sensitivity to outliers (in the presence of outliers, the models don't perform well)
 - high cost of computation

Hierarchical Clustering Dendrogram



<p>DBSCAN (Density-Based Spatial Clustering of Applications with Noise)</p>	<ul style="list-style-type: none"> density-based clustering mark data points far from each other as outliers then sort data based on commonalities <ul style="list-style-type: none"> MinPts: This is a certain number of neighbours or neighbour points Epsilon neighbourhood: This is a set of points that comprise a specific distance from an identified point. The distance between these points should be less than a specific number (epsilon). Core Point: This is a point in the density-based cluster with at least MinPts within the epsilon neighbourhood. Border point: This is a point in the density-based cluster with fewer than MinPts within the epsilon neighbourhood. Noise point: This is an outlier that doesn't fall in the category of a core point or border point. It's not part of any cluster. algorithm steps <ol style="list-style-type: none"> identify a core, the radius is given and create a group for each core point identify border points and assign them to their designated core points any other point that's not within the group of border points or core points is treated as a noise point advantages <ul style="list-style-type: none"> not require a specified number of clusters resourceful in the identification of outliers flexible in terms of the size and shape of clusters disadvantages <ul style="list-style-type: none"> not effective in clustering datasets that comprise varying densities fail to understand the data well may lead to difficulties in choosing a threshold core point radius in some rare cases, we can reach a border point by two clusters, which may create difficulties in determining the exact cluster for the border point 	 <p>DBSCAN Clustering in Iris Data</p> 
<p>EM (Expectation Maximum)</p>	<ul style="list-style-type: none"> a limitation of maximum likelihood estimation : it assumes that the dataset is complete, or fully observed an approach for performing maximum likelihood estimation in the presence of latent variables algorithm steps <ol style="list-style-type: none"> E-step : estimating the missing variables in the dataset M-step : then maximise the parameters of the model in the presence of the data, repeating these two steps until convergence 	 <p>Gaussian distribution</p> <p>Mixture</p>
<p>GMM (Gaussian Mixture Models)</p>	<ul style="list-style-type: none"> mixture model <ul style="list-style-type: none"> a model comprised of an unspecified combination of multiple probability distribution functions a statistical procedure or learning algorithm is used to estimate the parameters of the probability distributions to best fit the density of a given training dataset GMM <ul style="list-style-type: none"> is a mixture model that uses a combination of Gaussian (Normal) probability distributions requires the estimation of the mean and standard deviation parameters for each maximum likelihood estimate is the most commonly used technique for estimating the parameters 	<ul style="list-style-type: none"> example usage of GMM <ul style="list-style-type: none"> Consider the case where a dataset is comprised of many points by two different processes. The points for each process have distribution, but the data is combined and the distributions are obvious to which distribution a given point may belong. The processes used to generate the data point represents a latent variable and process 1. It influences the data but is not observable. the EM algorithm is an appropriate approach to use to estimate distributions. E-step : estimate a value for the process latent variable for each point M-step : optimise the parameters of the distribution using maximum likelihood estimation

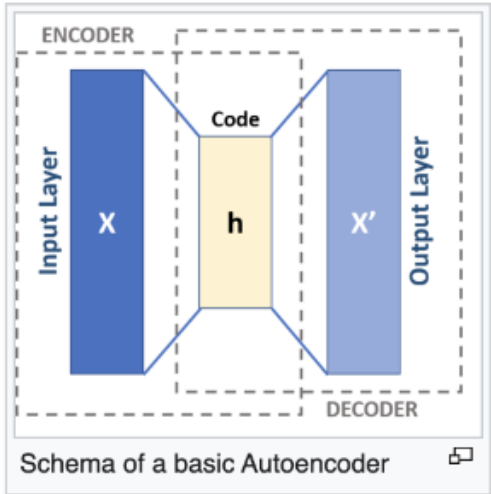
Clustering Time Series Data through Autoencoder-based Deep Learning Models

- Goal :** introduces a two-stage method for clustering time series data. First, a novel technique is introduced to utilise the characteristics (e.g., volatility) of given time series data in order to create labels and thus be able to transform the problem from unsupervised learning into supervised learning. Second, an autoencoder-based deep learning model is built to learn and model both known and hidden features of time series data along with their created labels to predict the labels of unseen time series data.

- Link : <https://arxiv.org/pdf/2004.07296.pdf> (04/2004)
- Code : in paper
- Credible source: -

autoencoder

- a type of neural networks that transforms input data into their output
- uses two parts in this transformation
 - encoder : transforms its high dimensional inputs into a smaller set of dimensions while keeping the most important features
 - decoder : the reduced set of features is used to reconstruct the initial input data
- latent-space representation
 - the output of the encoder
 - a compressed form of the input data in which the most influential and important features are kept



Experiment

training data	experiment	evaluation and output
Dataset	<div>1. first step : using k-means clustering, cluster no label time series data, and keep the cluster as label</div> <div>2. second step : using autoencoder based deep neural network, build prediction model</div> <div>it models hidden features and takes into account such features into the prediction</div>	<div><div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><</div></div>

- collect 70 companies listed by S&P 500 (https://en.wikipedia.org/wiki/List_of_S%26P_500_companies)
- scrap the time series data using read_html Python library
- capture adjusted close price from yahoo (01/Jan/2019 to 15/Apr/2019)

stock market data can be characterized

- volatility

■ It is the standard deviation of the changes in the values of a financial time series based on data

- of
te
n
u
s
e
d
to
d
e
m
o
n
st
r
at
e
th
e
ri
s
k
s
a
s
s
o
ci
at
e
d
w
it
h
st
o
c
k
in
di
c
es
- al
s
o
k
n
o
w
n
a
s
s
w
in
gs

■ r
ef
e
rs
to
th
e
d
e
g
r
e
e
of
v
a
ri
at
ion
of
a
tr
a
di
n
g
p
ri
c
e
s
e
ri
e
s
o
v
e
r
ti
me

■ is calculated by the standard deviation of logarithmic returns

■ shows the frequency and severity in which the market price of an investment fluctuates

- shows uncertainty of the future of the economic and financial series

- return
 - The return of a stock in a given time period can be def

in
e
a
s
th
e
n
at
u
r
al
lo
g
a
rit
h
m
of
th
e
cl
o
si
n
g
p
ri
c
e
at
th
e
e
n
d
of
th
e
p
e
ri
o
d
di
vi
d
e
d
b
y
th
e
cl
o
si
n
g
p
ri
c
e
of
th
e
st
o
c
k
at
th
e
e
n
d
of
th
e
p
r

e v i o u s p e r i o d		
----------------------------------------------------------	--	--

ref.

- <https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation>
- <https://developers.google.com/machine-learning/clustering/clustering-algorithms>
- <https://www.section.io/engineering-education/clustering-in-unsupervised-ml/>
- https://training.galaxyproject.org/training-material/topics/statistics/tutorials/clustering_machinelearning/tutorial.html
- <https://machinelearningmastery.com/expectation-maximization-em-algorithm/>