

Assignment Project Exam Help

Lecture 10: Clustering

GGR376

<https://powcoder.com>

Dr. Adams

Add WeChat powcoder

Clustering

Assignment Project Exam Help

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

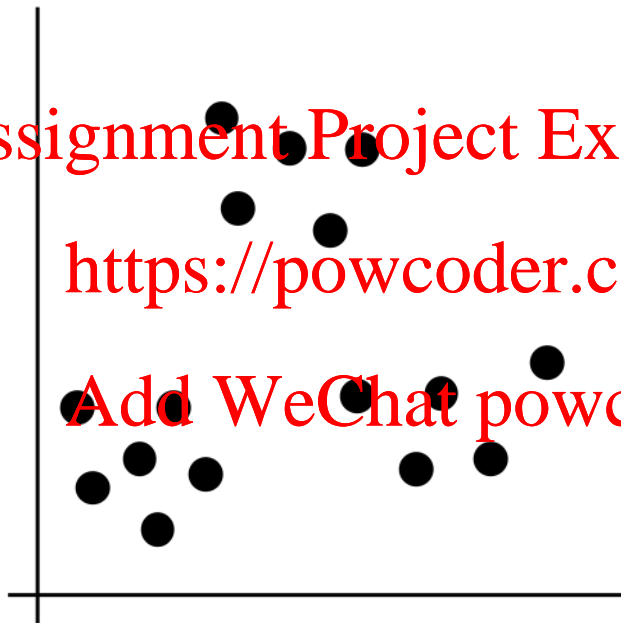
<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

<https://powcoder.com>

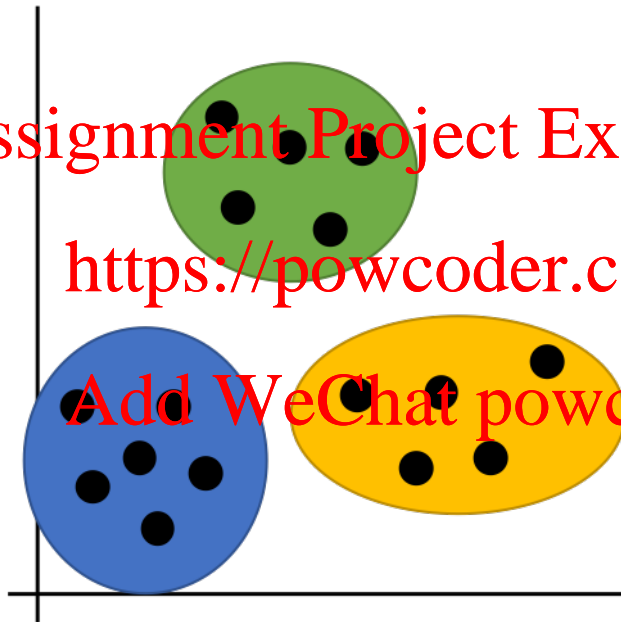
Add WeChat powcoder



Assignment Project Exam Help

<https://powcoder.com>

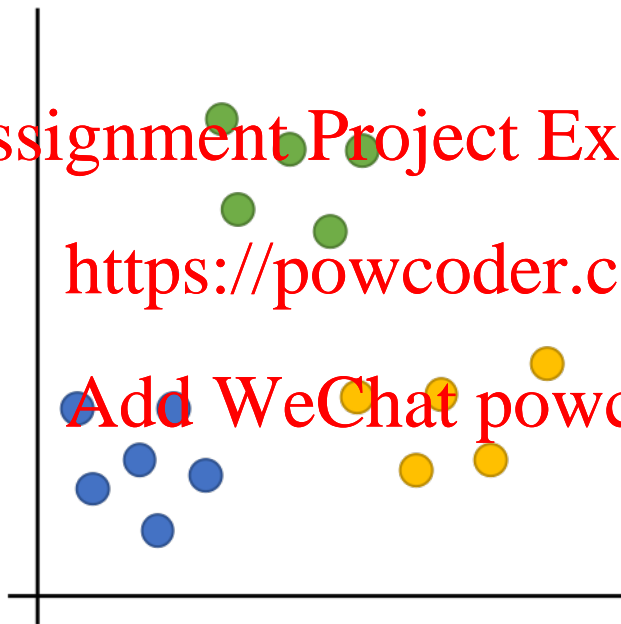
Add WeChat powcoder



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Iris Example



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

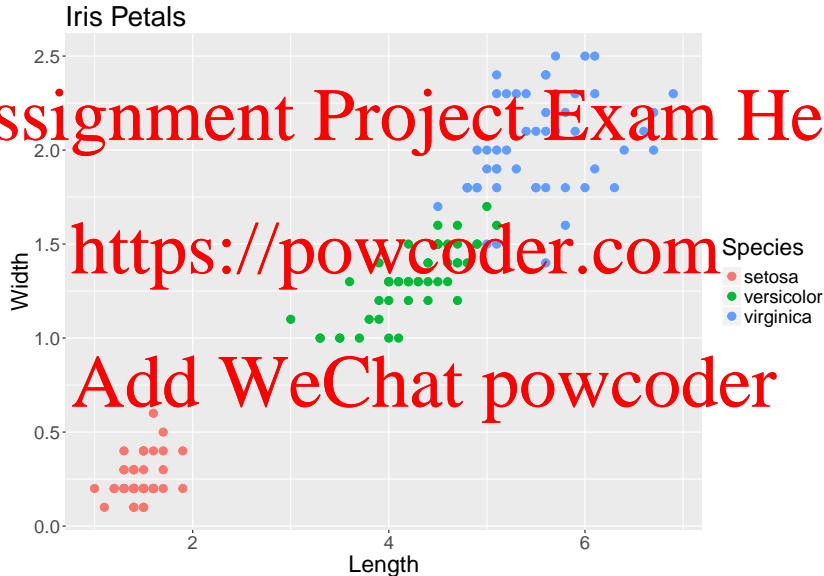
Example from: Teja Kodali Data Analyst at DV Trading LLC,
Chicago (IL)

Iris Data

```
library(datasets)
table(head(iris))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Petal Length and Width



Assignment Project Exam Help

```
set.seed(20)
irisCluster <- kmeans(iris[, 3:4], 3, nstart = 20)
# Cluster Sizes
irisCluster$size
```

<https://powcoder.com>

[1] 50 52 48

Add WeChat powcoder

Compare Clusters to Species

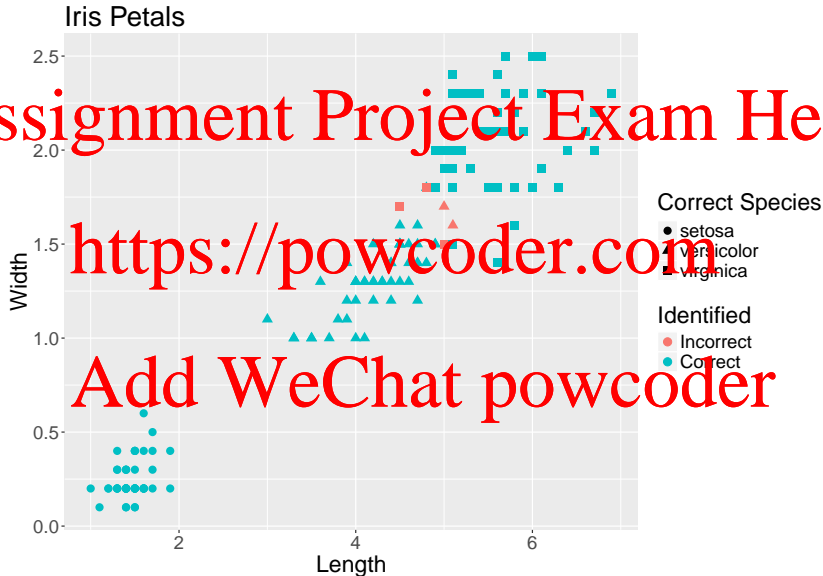
Assignment Project Exam Help

	setosa	versicolor	virginica
1	50	0	0
2	0	48	4
3	0	2	46

<https://powcoder.com>

Add WeChat powcoder

Plot Correct / Incorrect



Case Study Example

Assignment Project Exam Help

Kernel density estimation and K-means clustering to profile road accident hotspots

- ▶ “Every year approximately 3500 people are killed on Britain’s roads and 40,000 are seriously injured.”
- ▶ “Road traffic accidents according to the WHO are the leading injury related cause of death among people aged 10–24.”

Anderson, T. P. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention, 41(3), 359-364.

Assignment Project Exam Help

1. Present a methodology for the identification of high density accident zones using GIS and kernel density estimation (KDE)
2. Add attribute data to the accident zones.
3. Identify similar zones using a K-means clustering algorithm with regards to attribute data and compare.

<https://powcoder.com>
Add WeChat powcoder

Assignment Project Exam Help

- ▶ 5 years (1999–2003) of road accident data for the metropolitan area of London
- ▶ Kernel density calculated with traffic accidents
- ▶ Assigned attributes: Road length, Cycle lane length, Pedestrian crossings, London underground stations, Traffic lights, Bus stops, Schools, & Speed cameras.
- ▶ Applied k-means clustering to attribute data

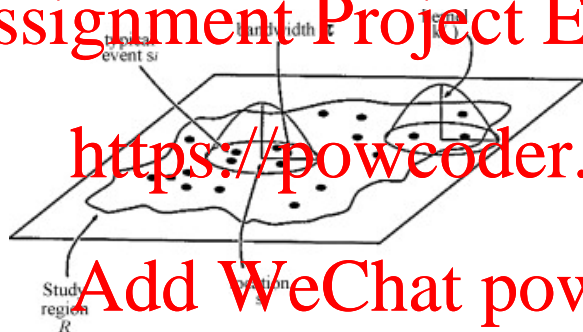
<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Assignment Project Exam Help

- ▶ Hot Spots were based on a threshold of accidents per cell.
- ▶ Convert point observations to a surface.

<https://powcoder.com>

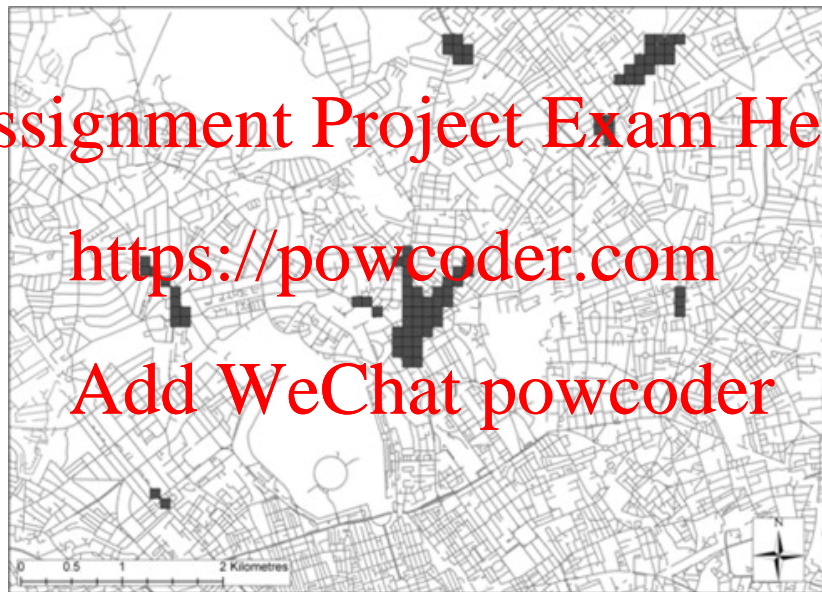
Add WeChat powcoder

Hot Spots

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Assignment Project Exam Help

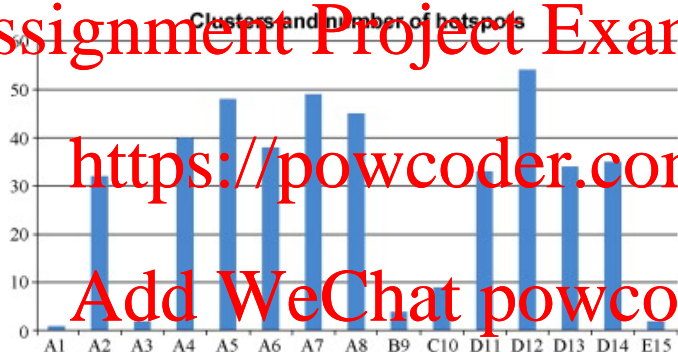
- ▶ Attempted to classify each of the 428 hotspots into relatively homogeneous types based on their environmental characteristics
- ▶ Given the number of hotspots identified in the kernel density
- ▶ K-means clustering was applied
 - ▶ K-means requires a specification of the number of clusters

Add WeChat powcoder

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Assignment Project Exam Help

- ▶ Extracted themes based on accident type and environmental conditions.
- ▶ Policy?

<https://powcoder.com>

Add WeChat powcoder

k -means Clustering

Assignment Project Exam Help

k -means clustering aims to partition n observations into k clusters, where each observation belongs to the cluster with the nearest mean, serving as a centre of the cluster.

<https://powcoder.com>

Add WeChat powcoder

k-means Overview

- ▶ Unsupervised clustering algorithm
 - ▶ You do not supply any information on clusters
- ▶ k is the number of clusters
 - ▶ Usually user specified
- ▶ An iterative approach
 - ▶ I.e. a solution is found through optimization
- ▶ Does not necessarily find a global solution
 - ▶ NP-Hard problem
 - ▶ Different solutions are possible each iteration
 - ▶ Especially as n gets large

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

k-means Video

Assignment Project Exam Help

K-means clustering: how it works

https://www.youtube.com/watch?v=_a/VzCGNrcf8

Add WeChat powcoder

Assignment Project Exam Help

- ▶ Locate k random centroids to initiate clusters.
- ▶ Compute for each n the distance to each centroid.
- ▶ Assign each n to nearest centroid.
- ▶ Compute the centroid mean

<https://powcoder.com>

- ▶ Iterate until convergence or you reach the max number of iterations.

Add WeChat powcoder

- ▶ Typically requires < 50 iterations.

Distance Measure

Assignment Project Exam Help

Euclidean Distance

Straight line distance between two points.

Depending on the data, you may want to standardize your values.

- ▶ If the units are drastically different

- ▶ Convert to a mean of 0 and a standard deviation of 1

Add WeChat powcoder

k -means challenges

Assignment Project Exam Help

1. Selecting k
2. There are actually k clusters.
3. SSE is the right objective to minimize.
4. Clusters have the same SSE.
5. All variables have the same importance for every cluster.

<https://powcoder.com>
Add WeChat powcoder

What if we don't know k

Assignment Project Exam Help

Not always - will you know how many clusters exist (or are required).

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

- ▶ k -means is a partitioning approach
 - ▶ The end result is a set of final clusters.
- ▶ Hierarchical methods
 - ▶ Divide data or aggregate in stages
 - ▶ End result includes all stages from a single cluster to n clusters
 - ▶ Select k clusters after the analysis

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

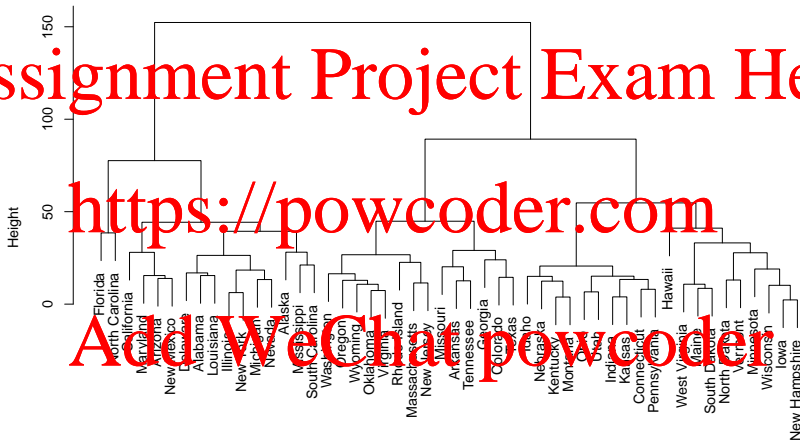
- ▶ Each point begins as its own cluster.
- ▶ Combine the two closest clusters
- ▶ Repeat until a single cluster remains.

<https://powcoder.com>

Add WeChat powcoder

Dendrogram

Cluster Dendrogram



Assignment Project Exam Help

<https://powcoder.com>

Acadwecm powcoder

Agglomeration methods (i.e, linkage methods)

Assignment Project Exam Help

- ▶ Complete Linkage
- ▶ Single Linkage
- ▶ Mean Linkage
- ▶ Centroid Linkage
- ▶ Ward's Method

<https://powcoder.com>

The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.

Add WeChat powcoder

Assignment Project Exam Help

- ▶ Computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2.
- ▶ Considers the largest value (i.e. maximum value) of these dissimilarities as the distance between the two clusters.
- ▶ Tends to produce compact clusters.

<https://powcoder.com>
Add WeChat powcoder

Assignment Project Exam Help

- ▶ Computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2.
- ▶ Considers the smallest of these dissimilarities as a linkage criterion.
- ▶ Tends to produce long, “loose” clusters.

<https://powcoder.com>
Add WeChat powcoder

Assignment Project Exam Help

- ▶ Computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2.
- ▶ Considers the average of these dissimilarities as the distance between the two clusters.

<https://powcoder.com>

Add WeChat powcoder

Centroid linkage clustering

Assignment Project Exam Help

It computes the dissimilarity between the centroid for cluster 1 and the centroid for cluster 2.

<https://powcoder.com>

Add WeChat powcoder

Ward's method

Assignment Project Exam Help

Minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

<https://powcoder.com>

Add WeChat powcoder

Spatially Constrained Clustering

Assignment Project Exam Help

Often times when dealing with spatial data we may have a reason to force data to be spatially contiguous.

Spatially constrained clustering is the approach.

- ▶ Regionalization
- ▶ Contiguity-constrained clustering

Add WeChat powcoder

<https://powcoder.com>

Gerry Mandering

Assignment Project Exam Help

<https://www.youtube.com/watch?v=Mhy11UJb9AY>
<https://powcoder.com>

Add WeChat powcoder

The Problem

Assignment Project Exam Help

The new area units may be at odds with the two objectives:

- ▶ Similarity of attributes
 - ▶ Non-spatial clustering
- ▶ Similarity of location
 - ▶ Spatial contiguity

<https://powcoder.com>

Add WeChat powcoder

These may not be aligned.

Assignment Project Exam Help

- ▶ Start with k -means result
- ▶ Split / combine the non-contiguous clusters
- ▶ The k number may change
- ▶ Inefficient approach

<https://powcoder.com>

Add WeChat powcoder

The Weighting Approach

Assignment Project Exam Help

Without explicit contiguity.

- ▶ Multi-objective
- ▶ Include x-y components as attributes in the clustering
 - ▶ Does not ensure contiguity
 - ▶ Choosing a weight for the x-y attributes is challenging
 - ▶ Eventually with strong enough weights

Add WeChat powcoder

<https://powcoder.com>

Automatic Zoning Procedure (AZP)

Assignment Project Exam Help

- ▶ Start with an initial solution
 - ▶ Output from k-mean
- ▶ Swapping polygons between zones
- ▶ No optimal solution
 - ▶ N-P hard problem

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

- ▶ Spatial contiguity is constructed with a graph

▶ Observations are nodes
▶ Connections are edges

<https://powcoder.com>

- ▶ Splitting edges to create clusters

Add WeChat powcoder

Assignment Project Exam Help

- ▶ Defined and an integer programming problem

- ▶ 1 or 0

- ▶ <https://powcoder.com>
- ▶ Specify a criteria to optimize

- ▶ e.g. maximum difference between groups

Add WeChat powcoder

Assignment Project Exam Help

- ▶ SKATER
- ▶ REDCAP
- ▶ Max-p

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

- ▶ R
 - ▶ spdep::skater
 - ▶ spdep::ClustGeo
 - ▶ Semi-constrained
- ▶ GeoDa
 - ▶ Open-source GUI
- ▶ RedCap
 - ▶ Open-source GUI

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

Spatially constrained clustering can be done outside of R, but you will need to describe how you did this in your methods.

<https://powcoder.com>

Add WeChat powcoder