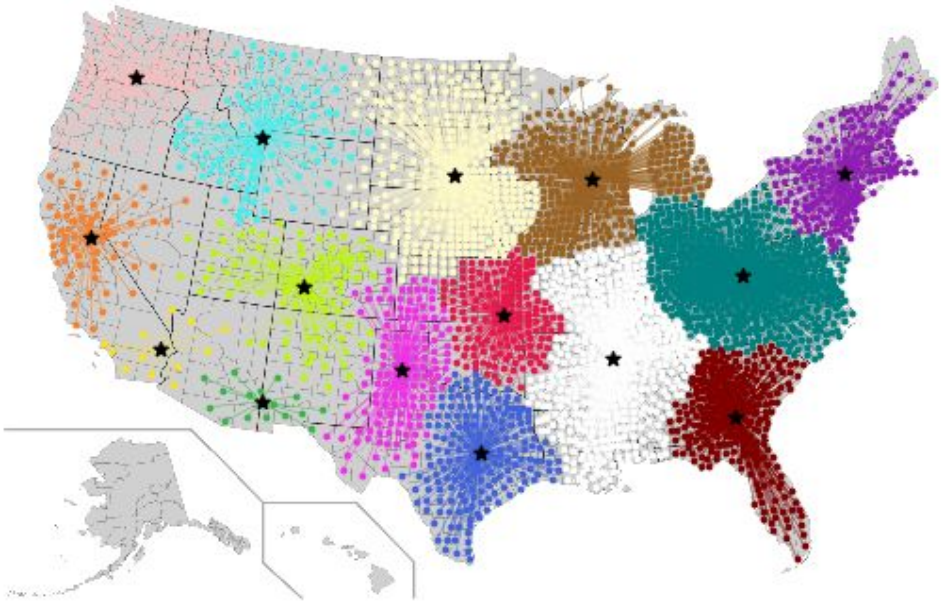
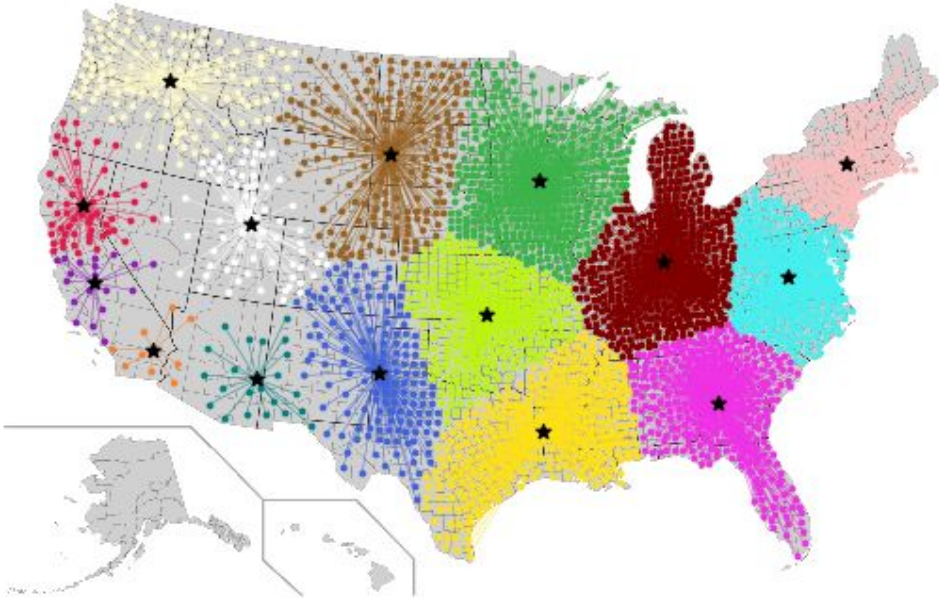
Laboratorio 4 - Clustering di dati medici

Alberto Bezzon, Tommaso Carraro, Alberto Gallinaro

Domanda 1



Domanda 2



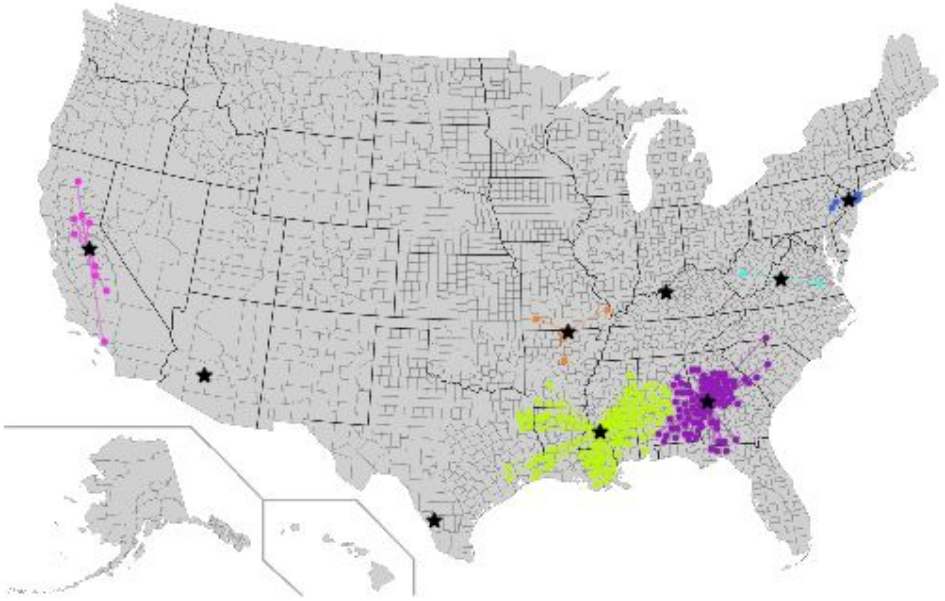
Domanda 3

Quando il numero di cluster di output è un numero piccolo o una piccola frazione del numero di punti del dataset il metodo di clustering k-means risulta più veloce rispetto al metodo di clustering gerarchico. Le complessità dei due metodi di clustering sono le seguenti:

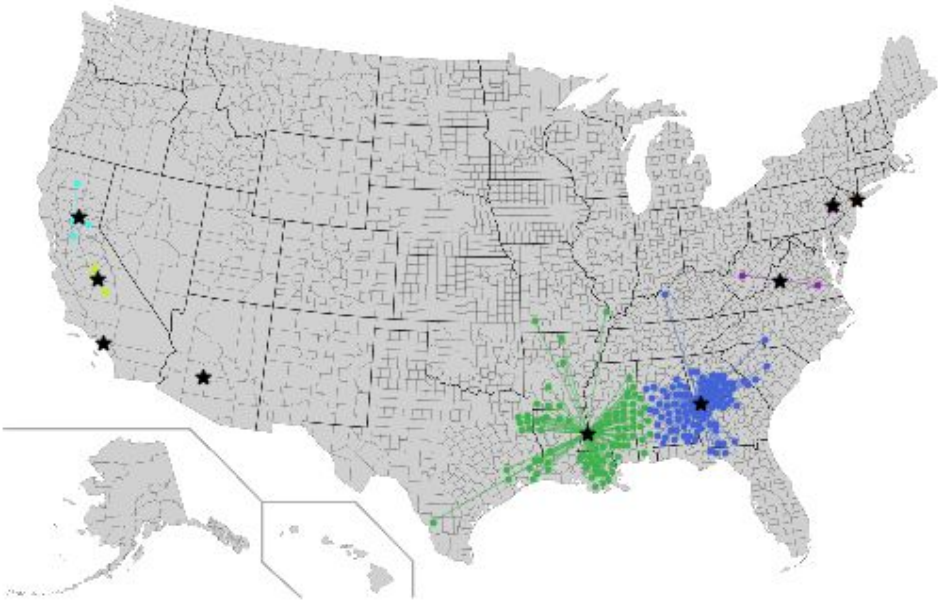
- k-means: O(qkn);
- gerarchico: O(n^2*log(n)), quando k è relativamente basso.

Poiché il numero di cluster in output è piccolo (15 cluster) e il numero di iterazioni anche (5 iterazioni), la complessità di k-means è praticamente O(n). Questo perché il numero di cluster e il numero di iterazioni sono trascurabili rispetto al numero di punti del dataset (3108 punti in questo caso).
Per quanto riguarda il clustering gerarchico, quando k è relativamente basso, come in questo caso (15 cluster), la sua complessità è O(n^2*log(n)) e questo spiega i tempi di gran lunga prolungati rispetto al clustering k-means.
Quando invece il numero di cluster si avvicina al numero di punti nel dataset, il clustering gerarchico ha complessità O(n*log(n)), mentre il clustering k-means ha complessità O(qn^2). In questo caso il clustering gerarchico sarebbe più veloce del clustering k-means.

Domanda 4



Domanda 5



Domanda 6

Distorsione clustering gerarchico: 196752213374.95956
Distrozione clustering k-means: 95382765365.33682

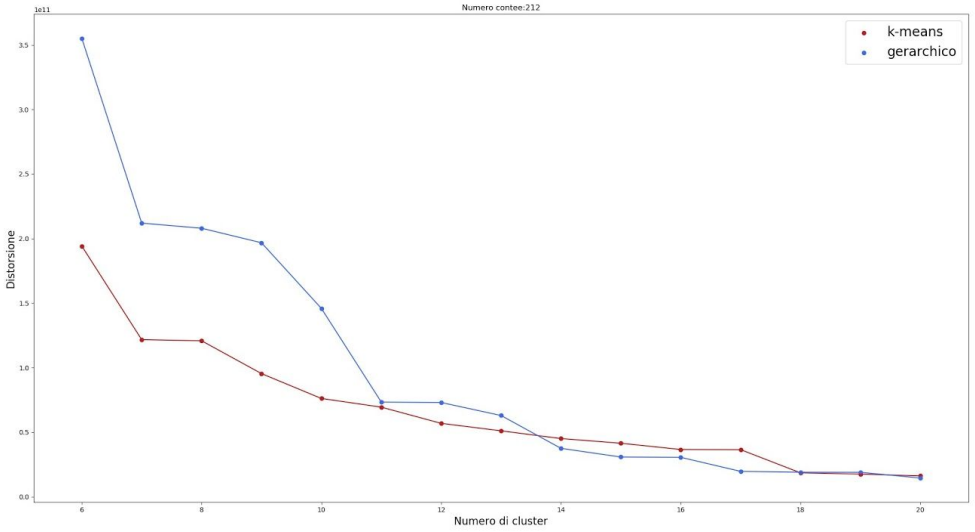
Domanda 7

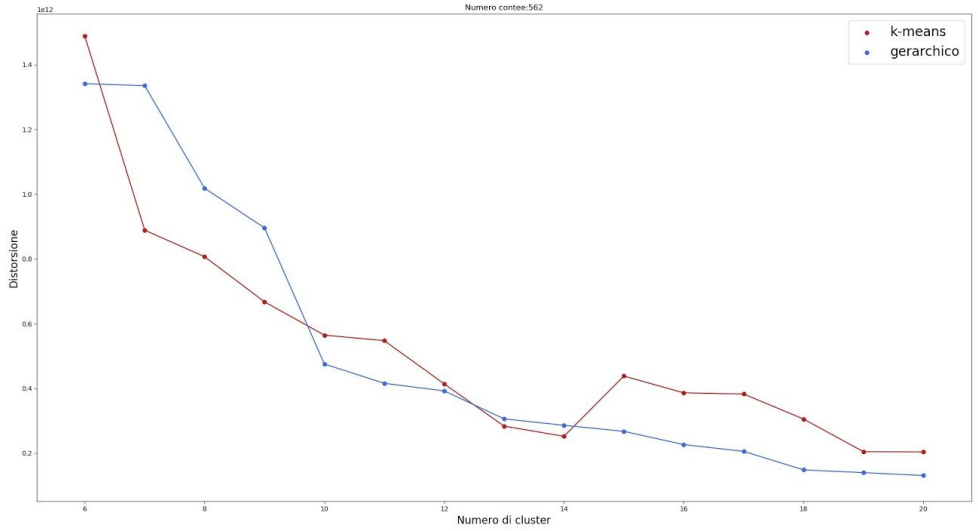
I due metodi di clustering si differenziano principalmente sui cluster che vengono generati nella costa occidentale degli Stati Uniti, mentre per quanto riguarda il resto delle contee i due metodi si comportano in maniera analoga.
Nella costa occidentale degli Stati Uniti il clustering gerarchico produce solamente due cluster, dove uno dei due ha un elevato errore mentre il secondo praticamente nullo.
Per quanto riguarda il clustering k-means, esso genera 4 cluster, di cui due con errore molto basso e due con errore praticamente nullo.
Per questo motivo il metodo di clustering gerarchico produce una distorsione più elevata rispetto al metodo di clustering k-means.
Questa disposizione dei cluster con basso errore nel metodo di clustering k-means è dovuta al fatto che tale metodo inizializza i cluster con le contee con il maggior numero di abitanti. Questa inizializzazione favorisce il metodo di clustering k-means al metodo di clustering gerarchico nel dataset con 212 contee.

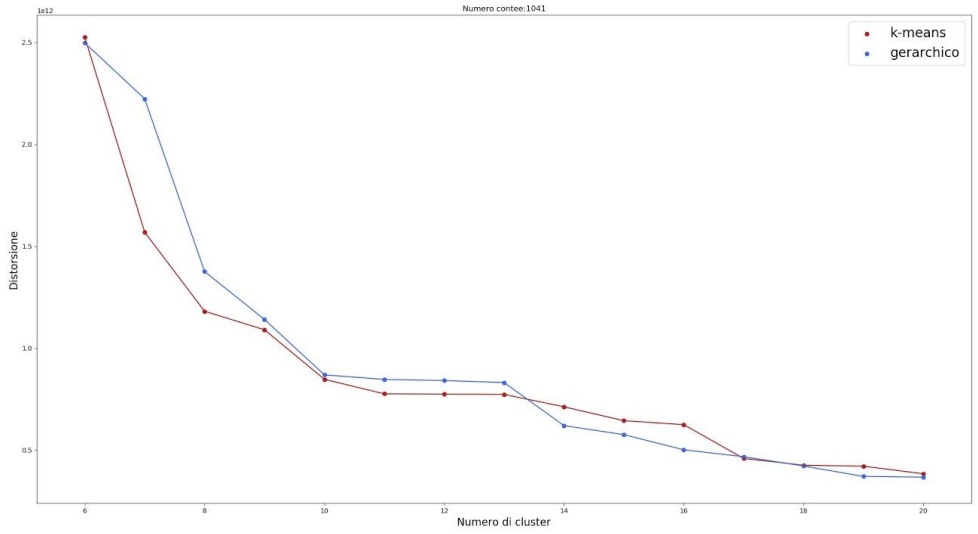
Domanda 8

Il metodo di clustering gerarchico richiede meno supervisione umana rispetto al metodo di clustering k-means. Infatti il metodo di clustering k-means richiede di fornire un numero di cluster desiderato, un numero di iterazioni massimo e i dati relativi alla popolazione nelle contee, nel caso di questo laboratorio.
Per quanto riguarda il clustering gerarchico, esso richiede solamente di scegliere il numero di cluster desiderato.

Domanda 9







Domanda 10

No, non esiste un metodo di clustering che produce sempre risultati con distorsione inferiore quando il numero di cluster di output è compreso tra 6 e 20. Si può però notare il seguente pattern nei grafici generati: sopra i 14 cluster l'algoritmo di clustering gerarchico ha distorsione sempre inferiore rispetto al metodo di clustering k-means.