



Progetto ML & Data Science

Andrea Vasciminno 904899

Vincenzo Cirillo 899870

Indice:

- » » » » » 1. DataSet Gas Turbine:
 - a. Analisi delle emissioni
 - b. Analisi di temperatura e umidità
 - c. Analisi temperatura di ingresso ed energia generata
- » » » » » 2. DataSet Apartment for rent
 - a. Clustering basato sulle similarità
 - b. Clustering geografico basato su similarità
 - c. Clustering geografico



Gas Turbine

Il dataset contiene al suo interno circa 35 mila istanze relative all'analisi dei dati degli impianti delle turbine a gas presenti nel territorio turco. Vengono riportate informazioni suddivise in 5 anni, dal 2011 al 2015.

L'analisi sui dati è stata svolta nella seguente maniera:

- Analisi delle emissioni
- Analisi di temperatura ed umidità
- Analisi della temperatura in ingresso e dell'energia generata

Gli algoritmi di clustering selezionati sono stati K-means e DB-Scan, il secondo è stato scelto dopo aver analizzato la numerosa presenza di outliers per garantire un'analisi meno viziata dagli stessi

Analisi delle emissioni

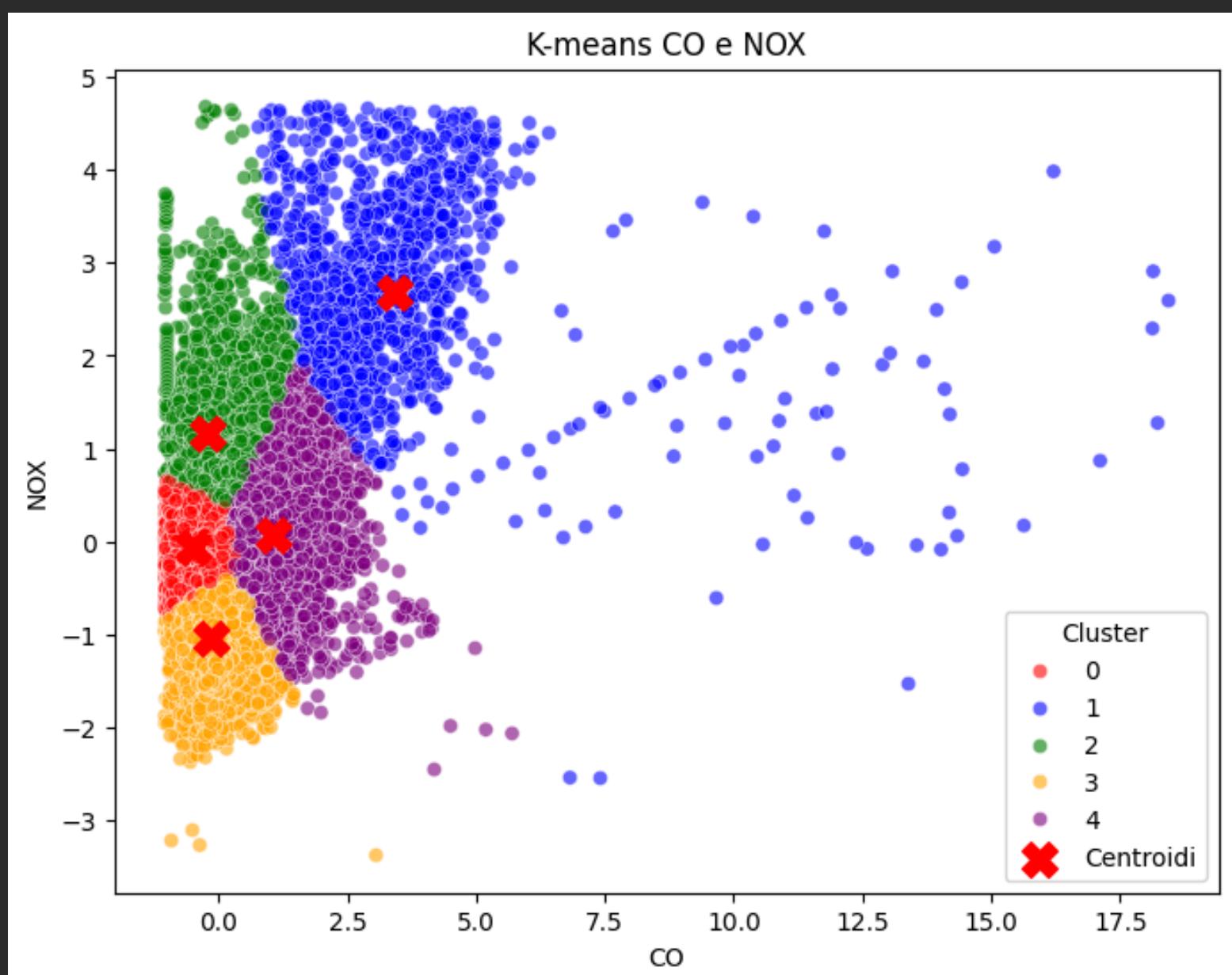
La prima analisi riguarda le emissioni di CO (Ossido di Carbonio) e di NOX (Ossido di Azoto), così da poter analizzare quante e quali sostanze vengono prodotte in maggioranza dalla combustione

K-means

L'elbow method evidenzia 5 cluster.

Si evidenzia:

- Un livello di combustione generalmente buono (bassa CO)
- I cluster giallo e rosso hanno emissioni ottimali
- il cluster blu mantiene un buon livello di CO, ma ha dei livelli di NOX alti, potrebbe essere causato o da temperature sopra la media o da situazioni particolari



Analisi delle emissioni

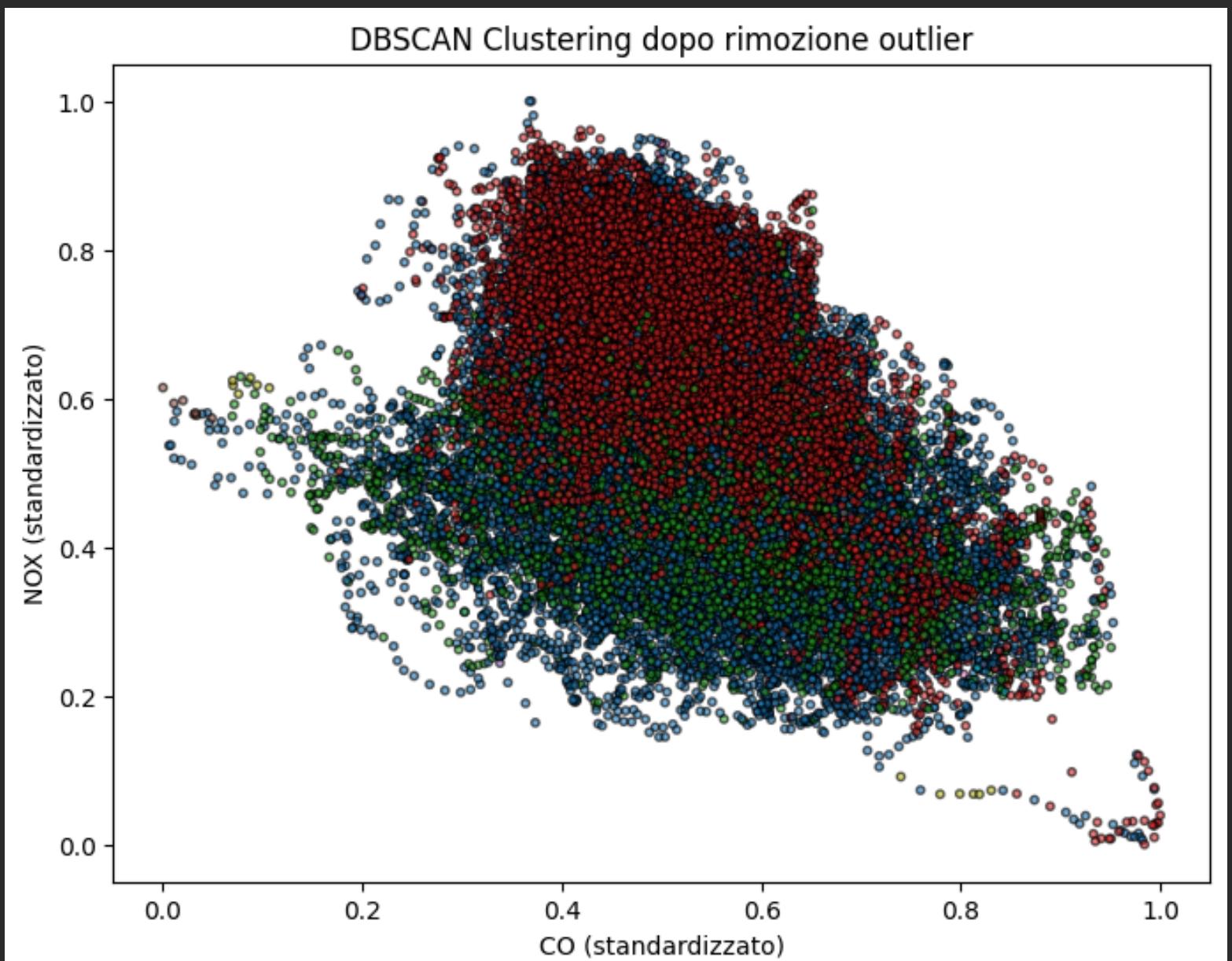
DB-Scan

Dal DB-Scan è possibile notare come eliminati gli outlier la divisione non risulti più così netta come nel K-means.

I dati sono standardizzati in maniera differente per garantirne migliore fruibilità.

Evidenziamo:

- Emissioni generalmente nella media
- I cluster esterni potrebbero essere soggetti a situazioni particolari o condizioni atmosferiche più rigide



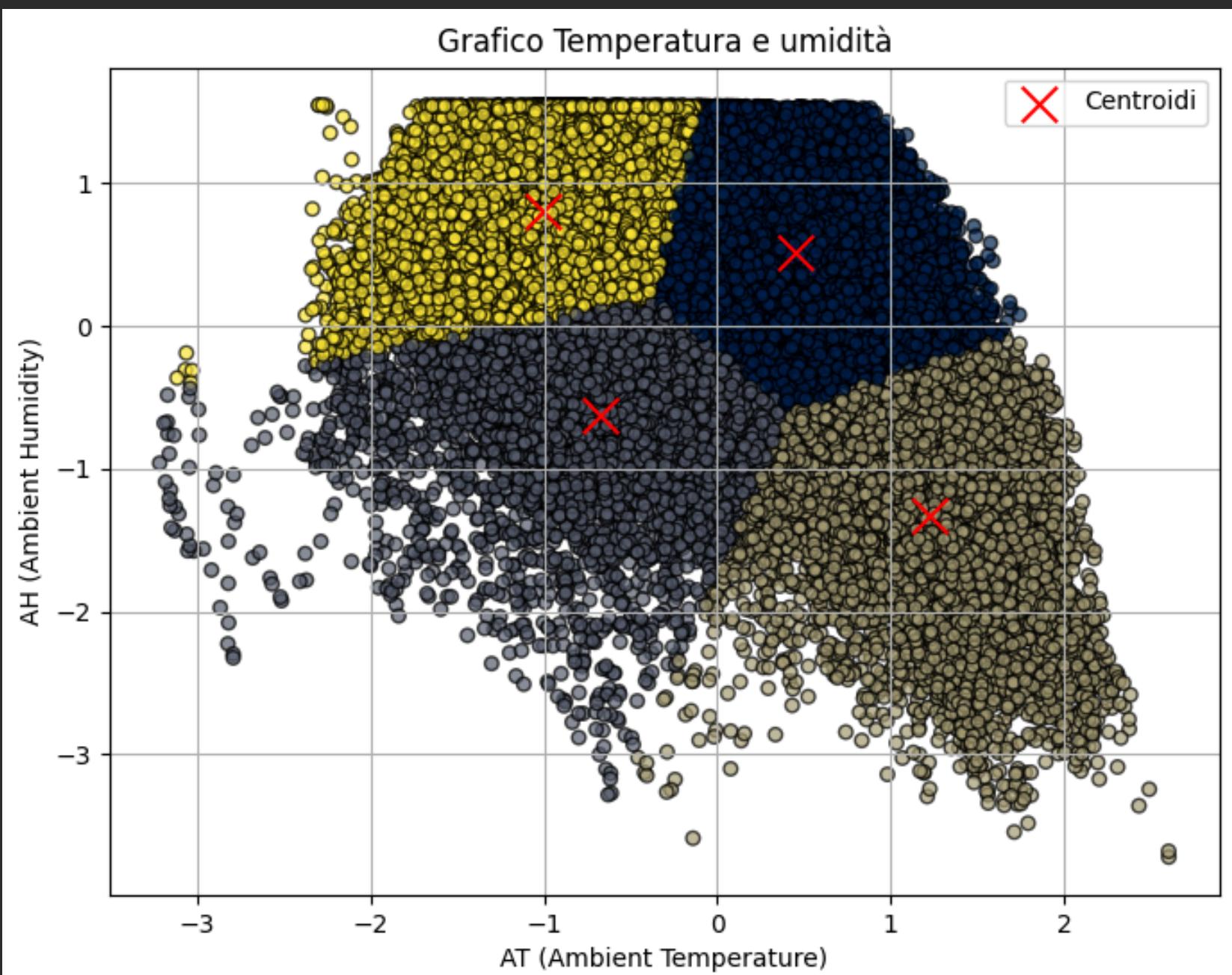
Analisi di temperatura e umidità

Effettuiamo un'analisi su temperatura ed umidità ambientali che impattano in maniera diretta sulle emissioni e sulle prestazioni delle turbine

K-means

L'elbow method evidenzia 4 cluster.

- I cluster centrali, vicini ai parametri (0,0), rappresentano le condizioni ottimali
- Nei punti che evidenziano un valore negativo di temperatura ma positivo di umidità le emissioni di CO potrebbero essere più alte



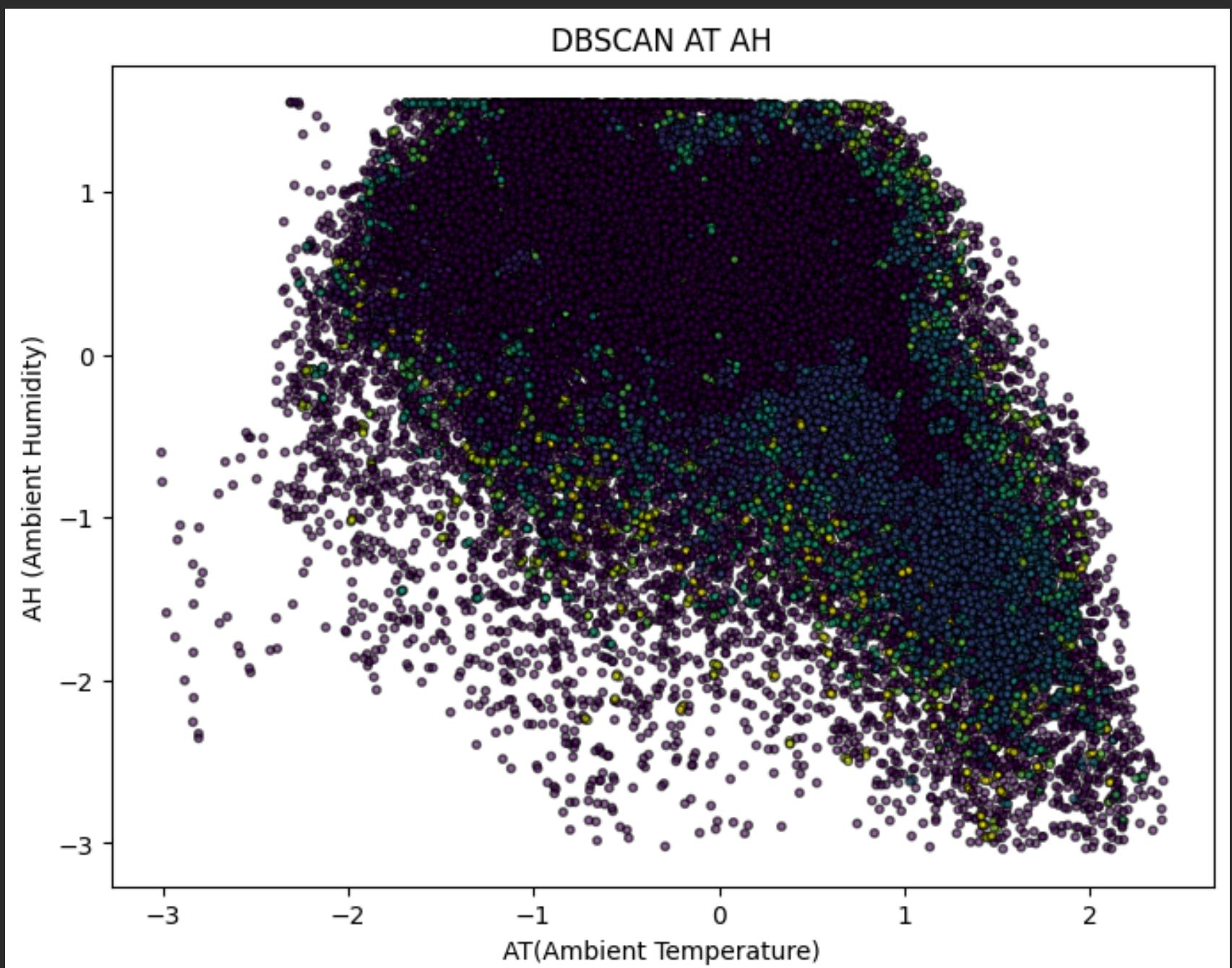
Analisi di temperatura e umidità

DB-Scan

Il DB-Scan evidenzia un andamento comune per la grande maggioranza dei dati.

Considerando un'umidità bassa aumenta la dispersione dei punti indicando indecisione in questa specifica condizione,

La distribuzione triangolare suggerisce come la variazione dell'umidità sia dettata dalla variazione della temperatura secondo tendenze specifiche.



Analisi temperatura ingresso ed energia generata

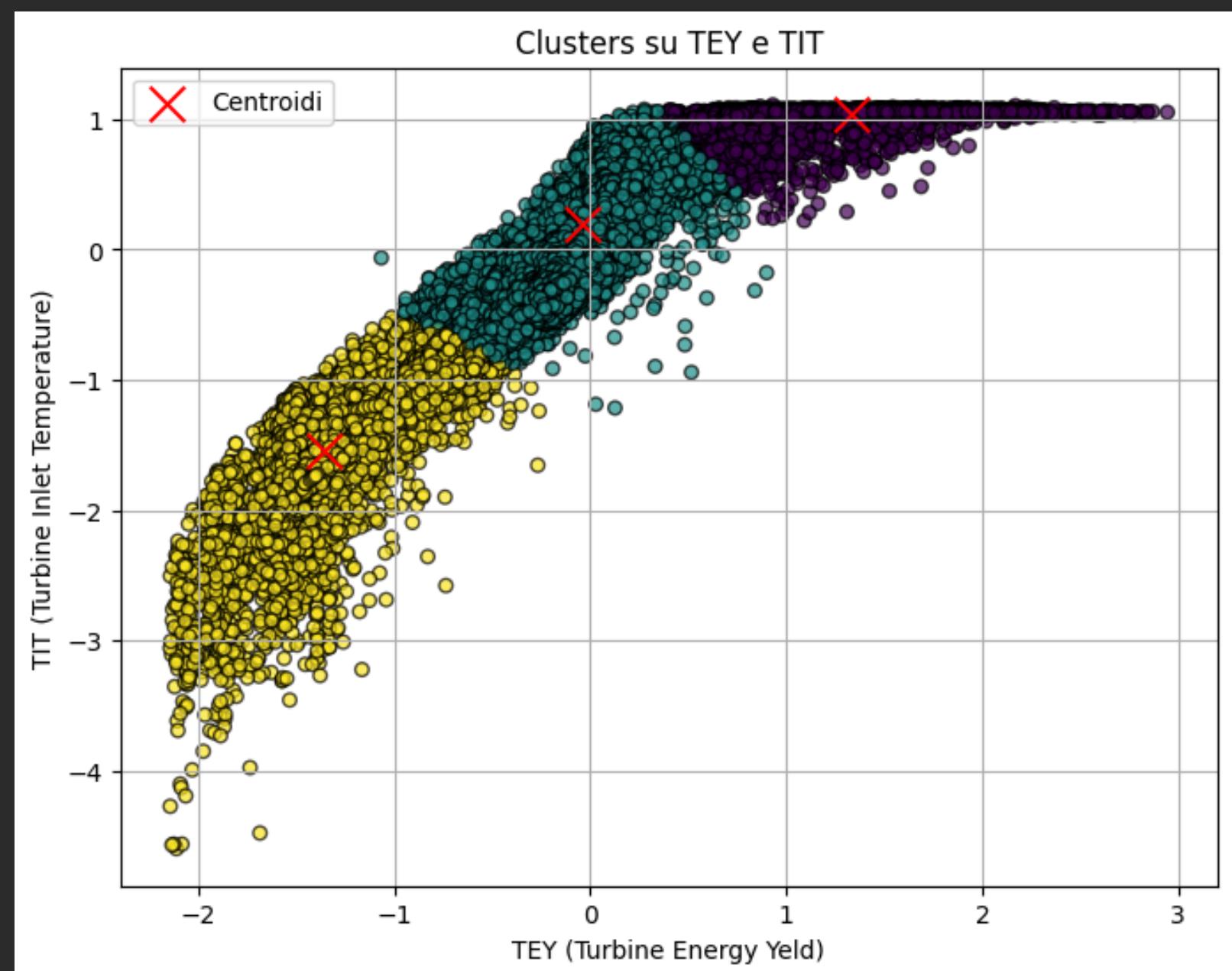
Analizzando la temperatura d'ingresso e l'energia generata dalla turbina è possibile comprendere come la produzione energetica viene influenzata. Inoltre l'analisi risulta sempre significativa anche per essere rapportata ai valori precedenti

K-means

L'elbow method evidenzia 3 cluster, si può notare come i due valori sembrino direttamente proporzionali tra loro.

Considerazioni:

- il cluster giallo potrebbe indicare turbine che lavorano a bassa efficienza o in condizioni di potenza ridotta
- Il cluster verde sembra indicare uno stato di lavoro normale mentre il viola ad alto rendimento



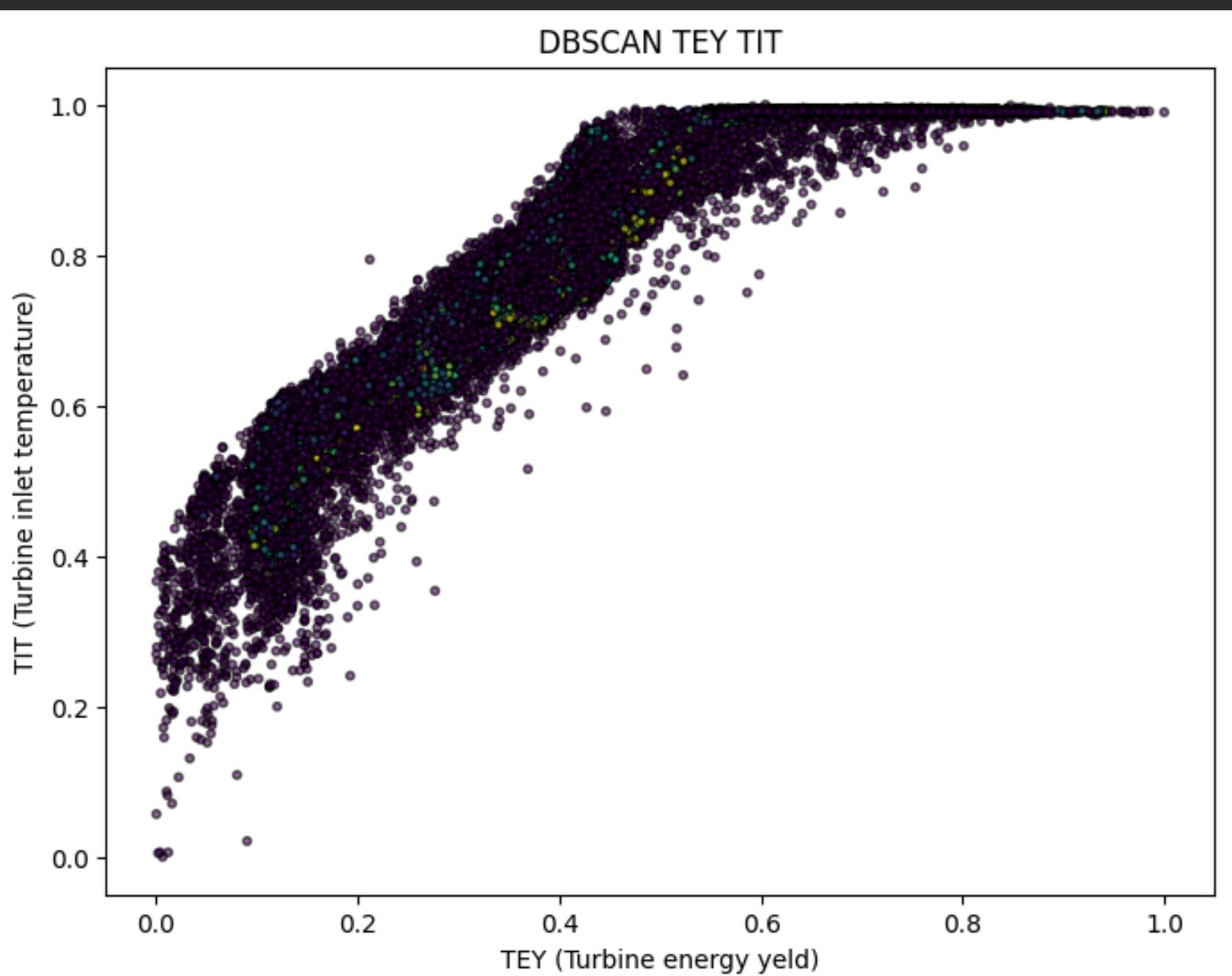
Analisi temperatura ingresso ed energia generata

DB-Scan

Anche il DB-Scan evidenzia una forte correlazione tra i due parametri.

Osservazioni:

- I cluster più densi potrebbero rappresentare le condizioni normali delle turbine
- Identificare gli outlier potrebbe essere utile per identificare impianti a cui effettuare manutenzione



Apartment rent dataset

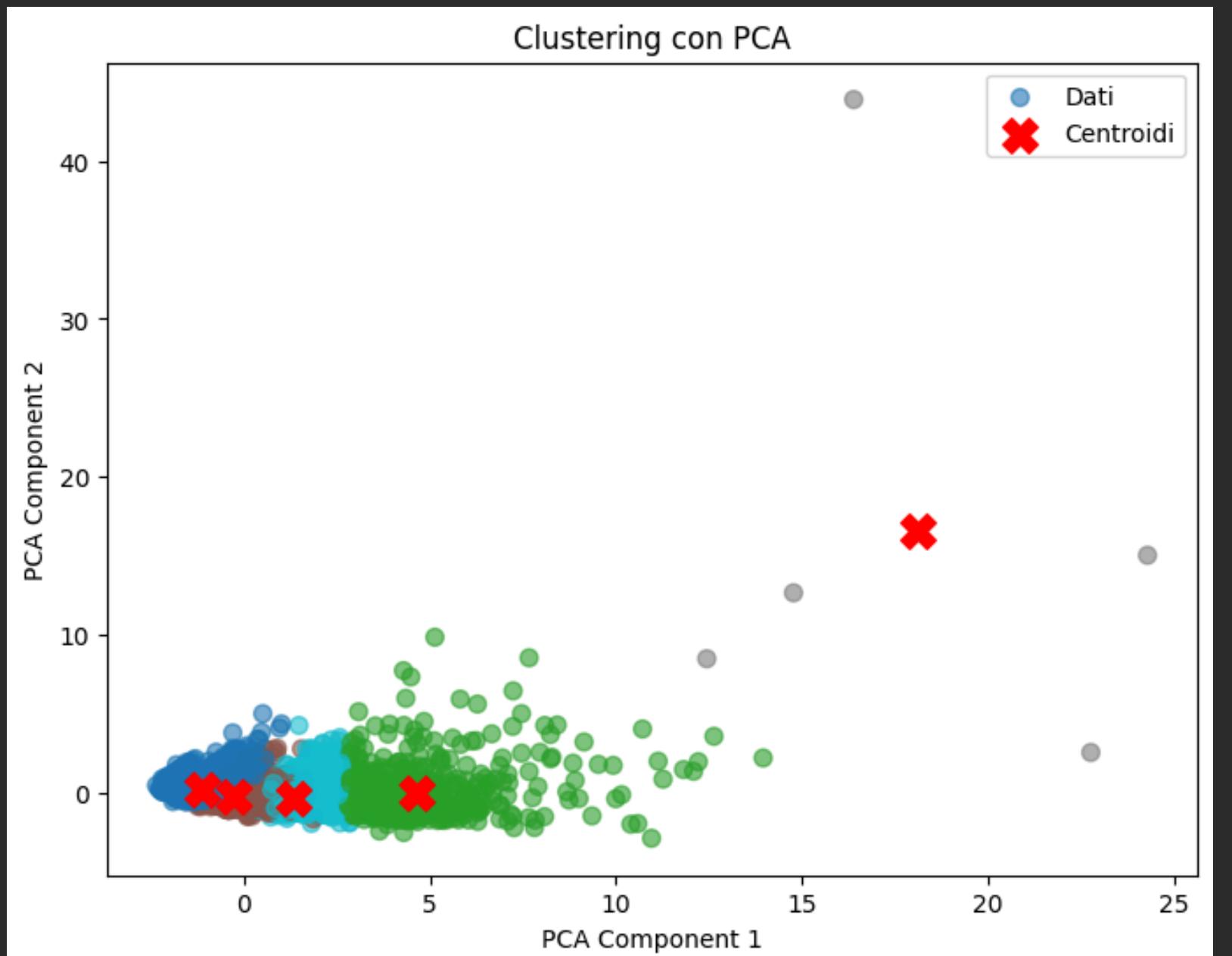
Il dataset contiene al suo interno circa 10 mila istanze di case in affitto sul territorio degli Stati Uniti d'America
Il dataset è stato lavorato nella seguente maniera:

- Drop dei valori Nand, per rimuovere istanze inutili al clustering
- Per i 3 clustering sono stati scelti gruppi di dati differenti:
 - “Price” “Square feet” “Bedrooms” “Bathrooms”
 - “Price” “Square feet”
 - “Longitude” “Latitude”
- I parametri scelti sono solo di tipo numerico, poichè il K-Means supporta solo questo tipo di parametri

Clustering basato sulla similarità

K-Means

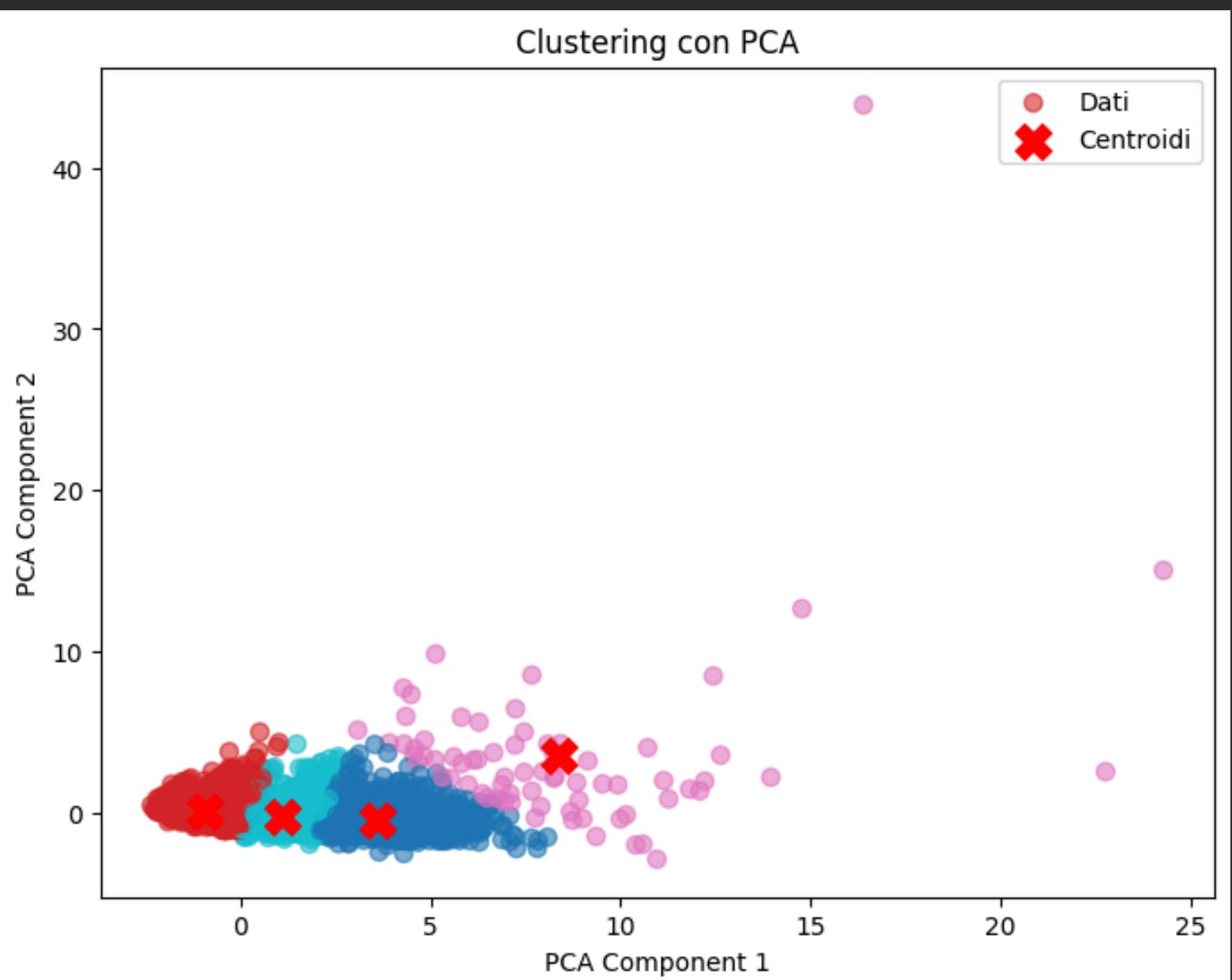
L'utilizzo del metodo del gomito per stilare il numero di cluster ottimale fornisce 5 come soluzione ottimale. Analizzando il grafico prodotto, però, si nota come il numero di cluster possa essere sbagliato. Questo per via della troppa densità di alcuni cluster e per centroidi posizionati in presenza di outlier. Per questo motivo si è scelto di decrementare il numero di cluster.



Clustering basato sulla similarità

K-Means

Riducendo il numero di cluster K-Means genera 4 cluster. I cluster sono ben definiti, distinti ed omogenei. Vengono però inseriti, come da aspettative, tutti gli outlier, che potrebbero rappresentare ad esempio case di lusso a prezzi anomali

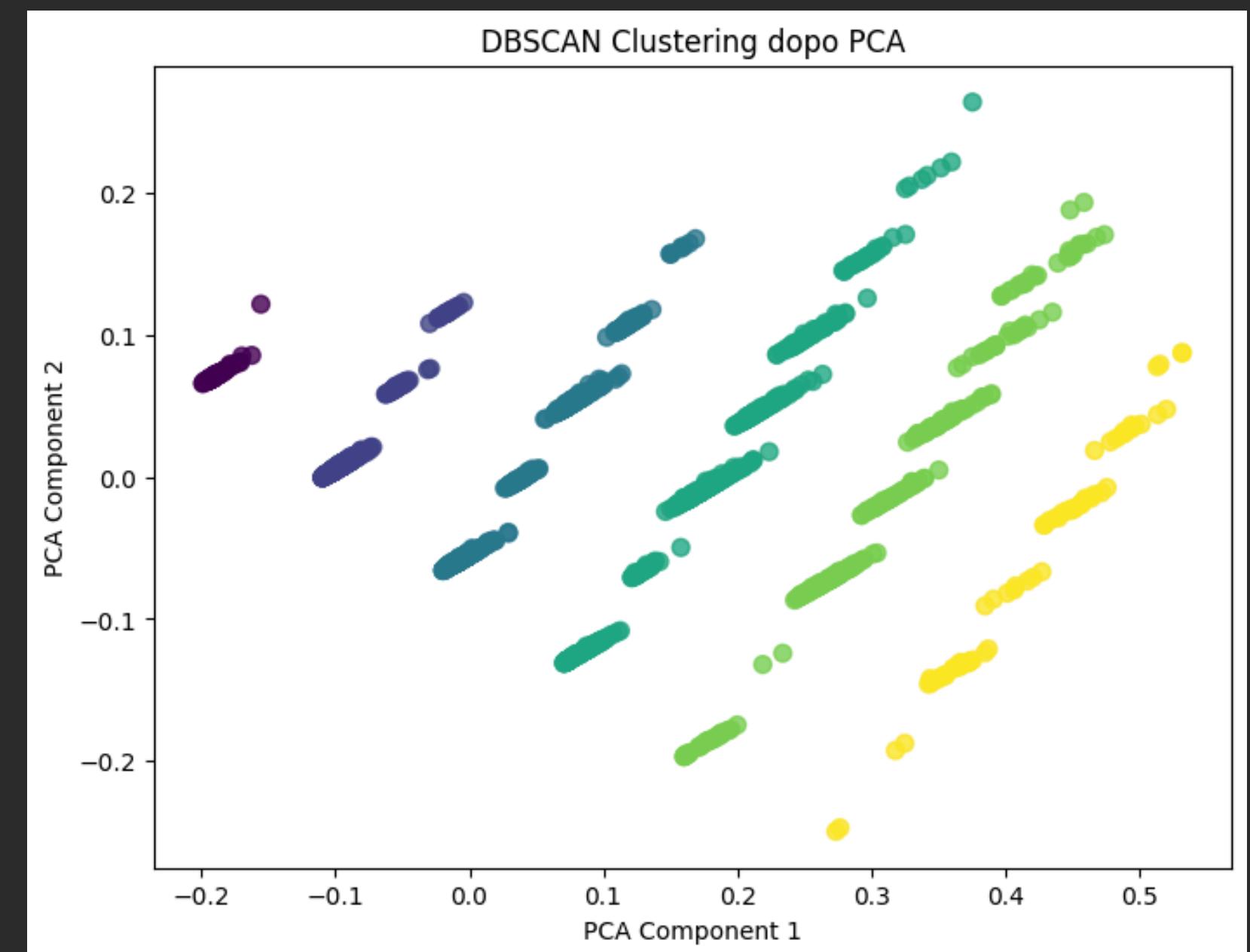


Clustering basato sulla similarità

DB-Scan

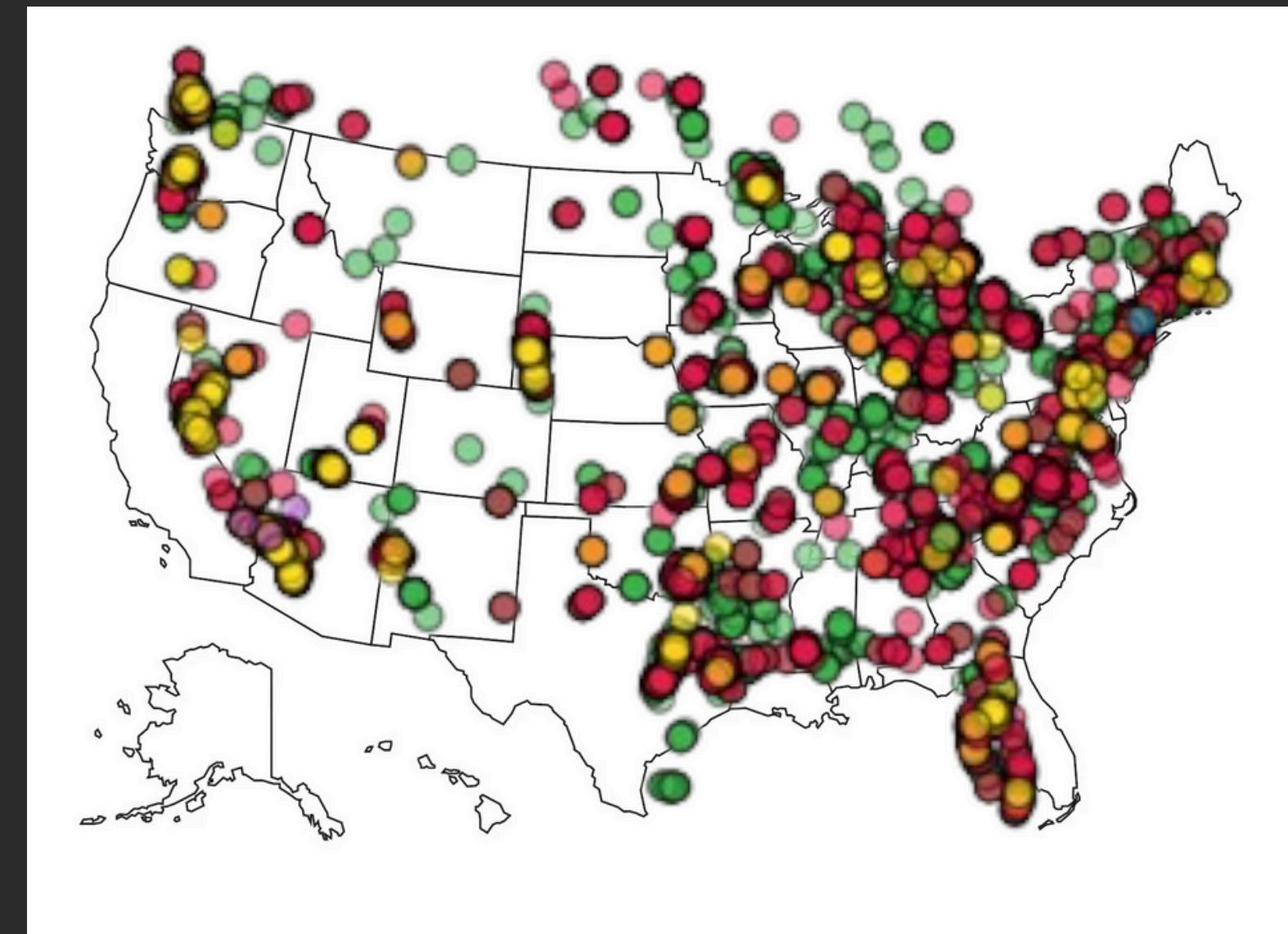
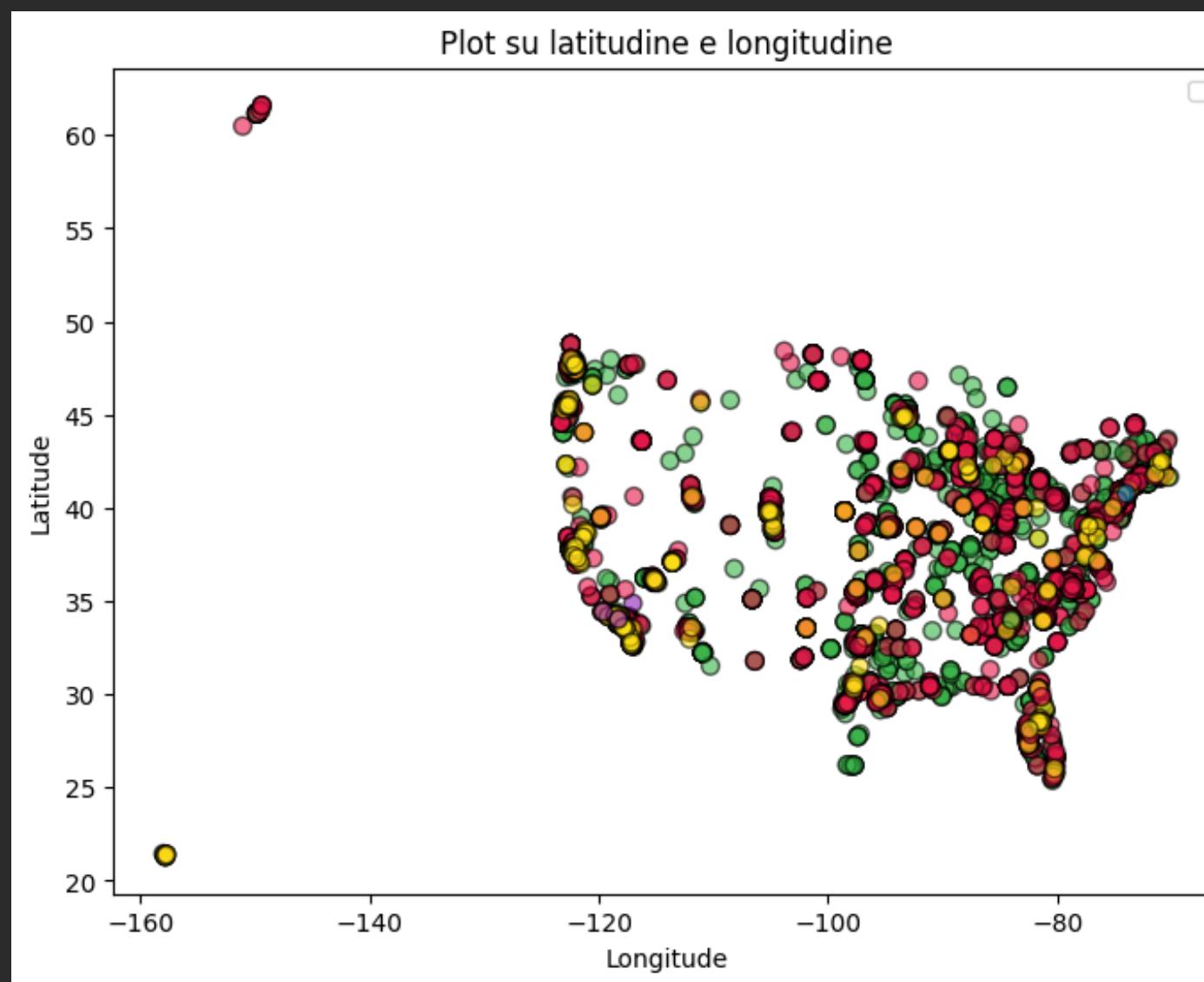
DB-Scan genera più cluster, dividendo le case secondo gli stessi parametri precedenti

Ottenendone di più, si genera maggiore precisione sui differenti tipi di abitazioni. Inoltre rimuove dalla visualizzazione i dati outliers visualizzabili nell'ultimo cluster K-Means. La rimozione degli outlier potrebbe aver diminuito la densità dei cluster.



Clustering geografico basato su similarità

K-Means



Clustering geografico basato su similarità

K-Means

K-Means normalizza i dati e, successivamente, genera 5 cluster dividendo le case in base a:

- Prezzo
- Metri quadri

In questo modo si riescono a creare cluster contenenti tutte le case che hanno grandezza e prezzo simili.

Il plot viene poi fatto su latitudine e longitudine in modo da poterle poi collocare graficamente sulla cartina degli Stati Uniti.

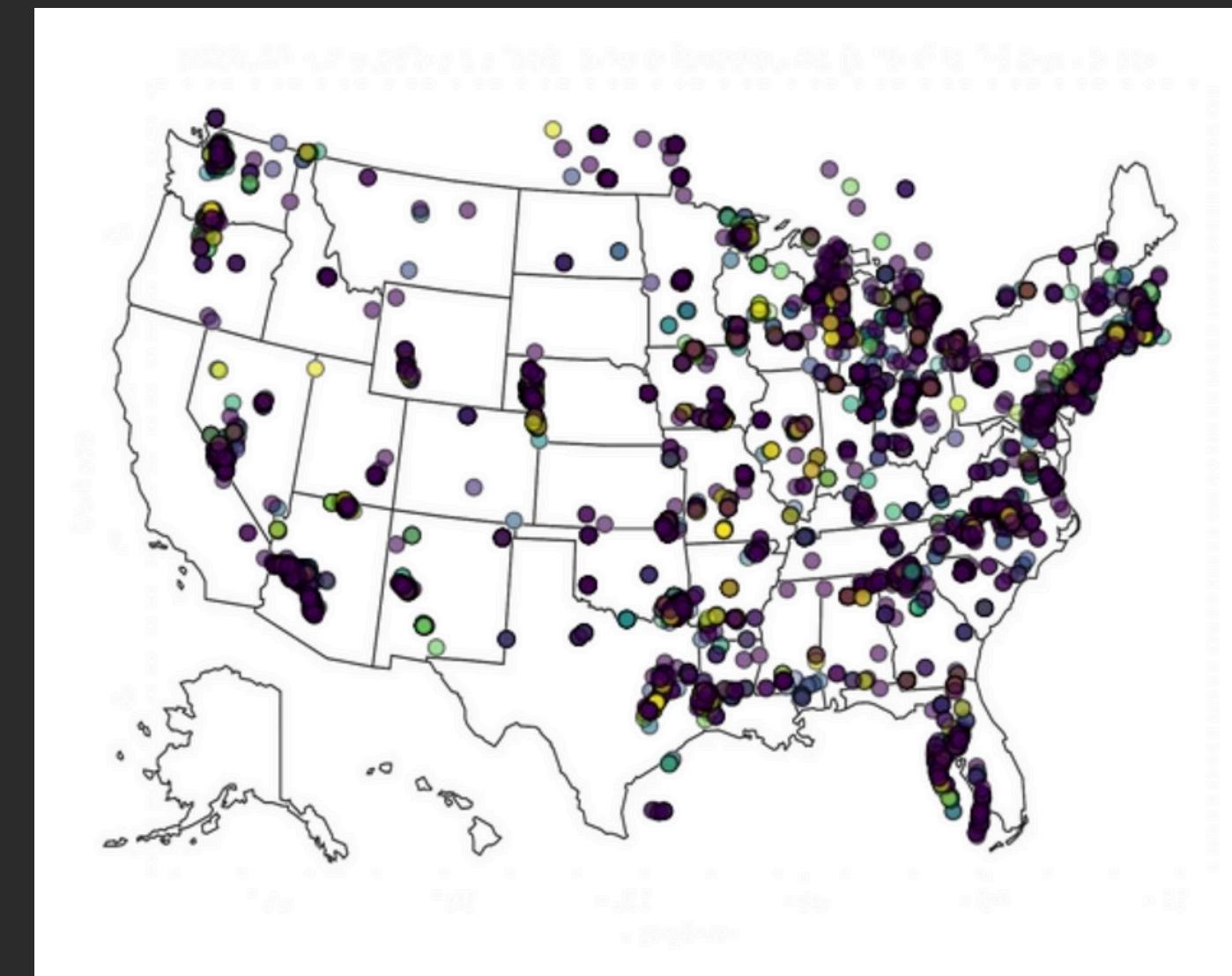
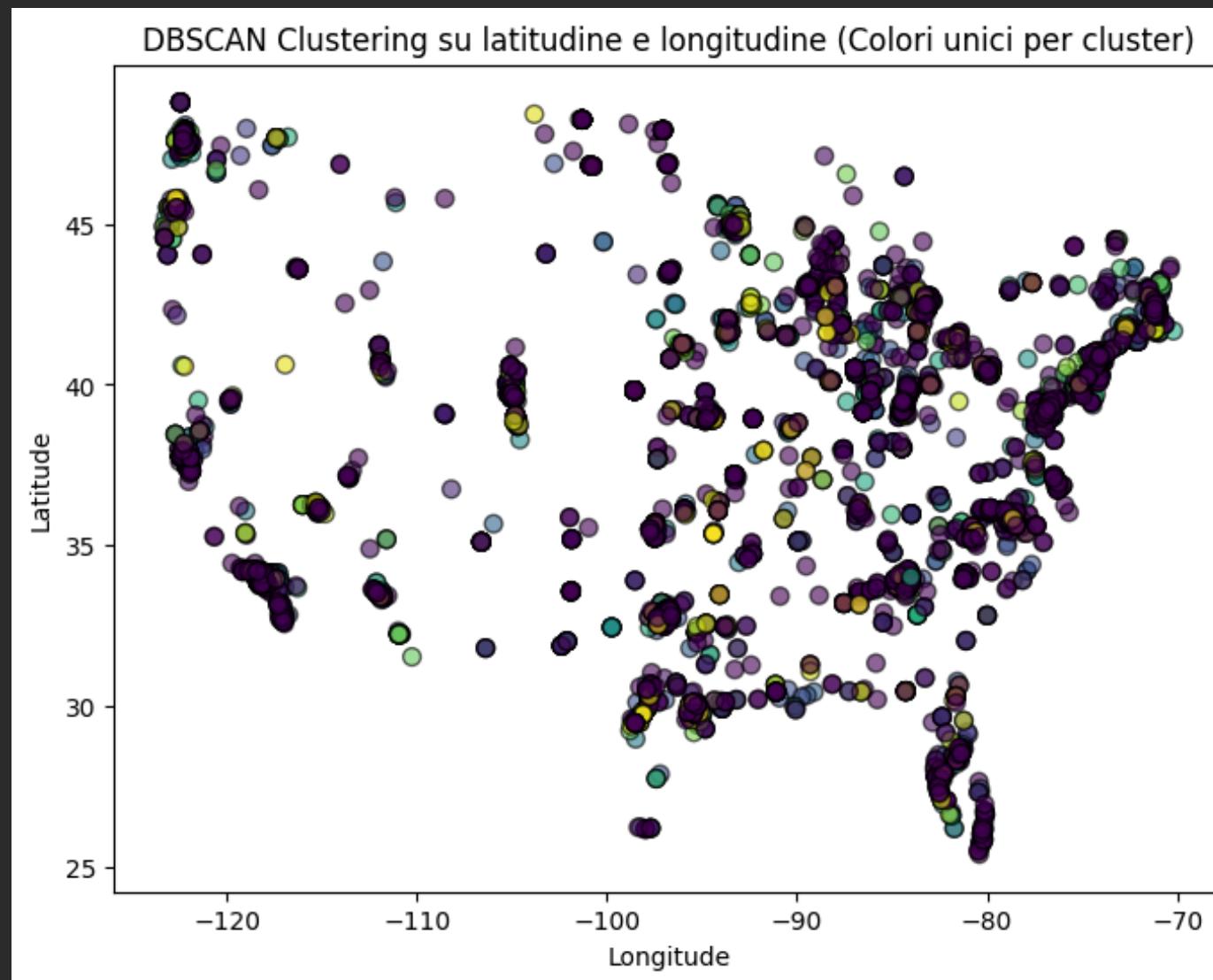
Analisi risultati:

La funzione di normalizzazione dei dati utilizzata storce leggermente le proporzioni terrestri.

In ogni caso si riescono a collocare sul territorio case divise in cluster di prezzi e metri quadri. Si ottiene una visione abbastanza generale. Questo permette di effettuare un'analisi globale del mercato in maniera abbastanza rapida

Clustering geografico basato su similarità

DB-Scan



Clustering geografico basato su similarità

DB-Scan

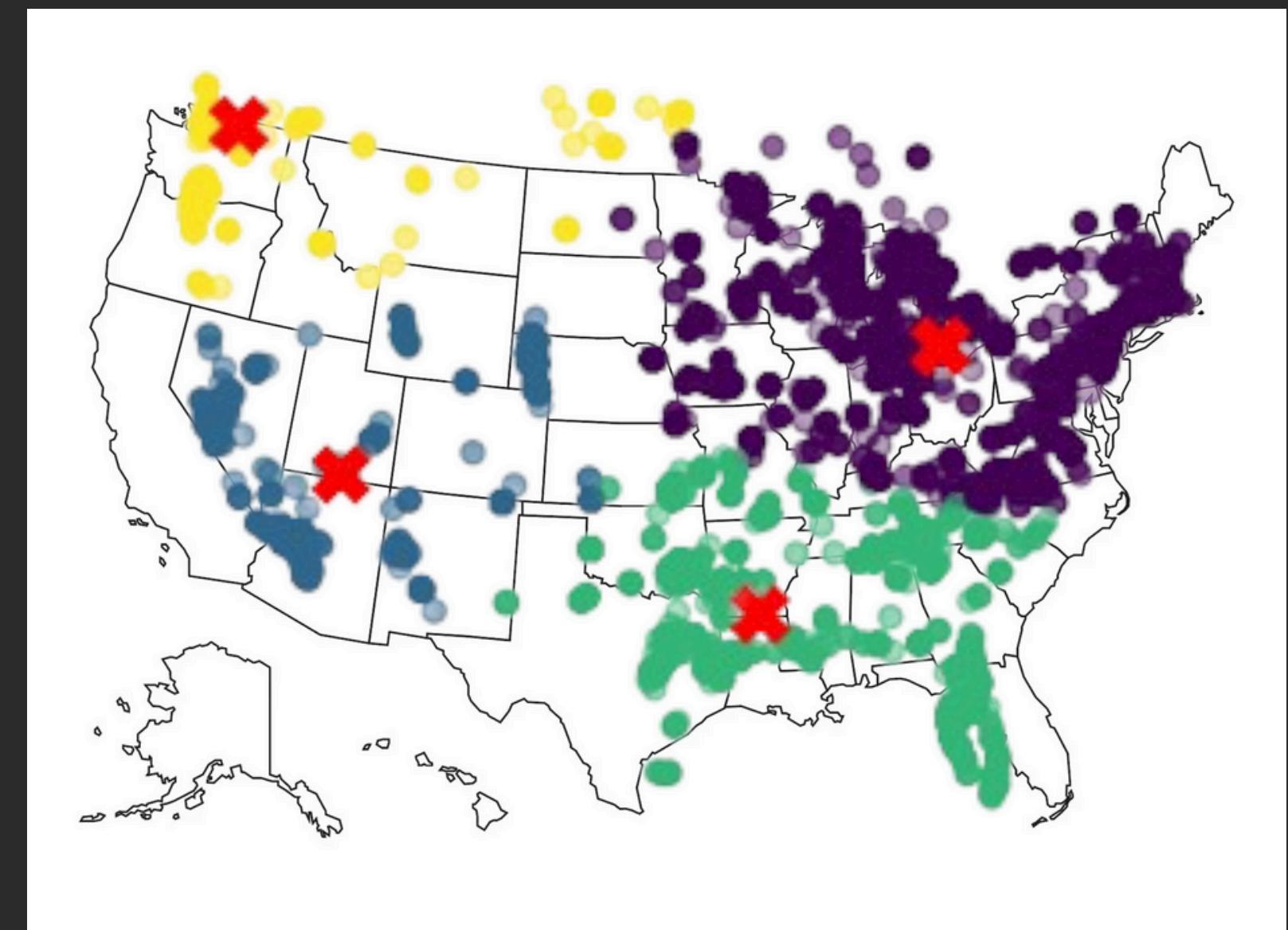
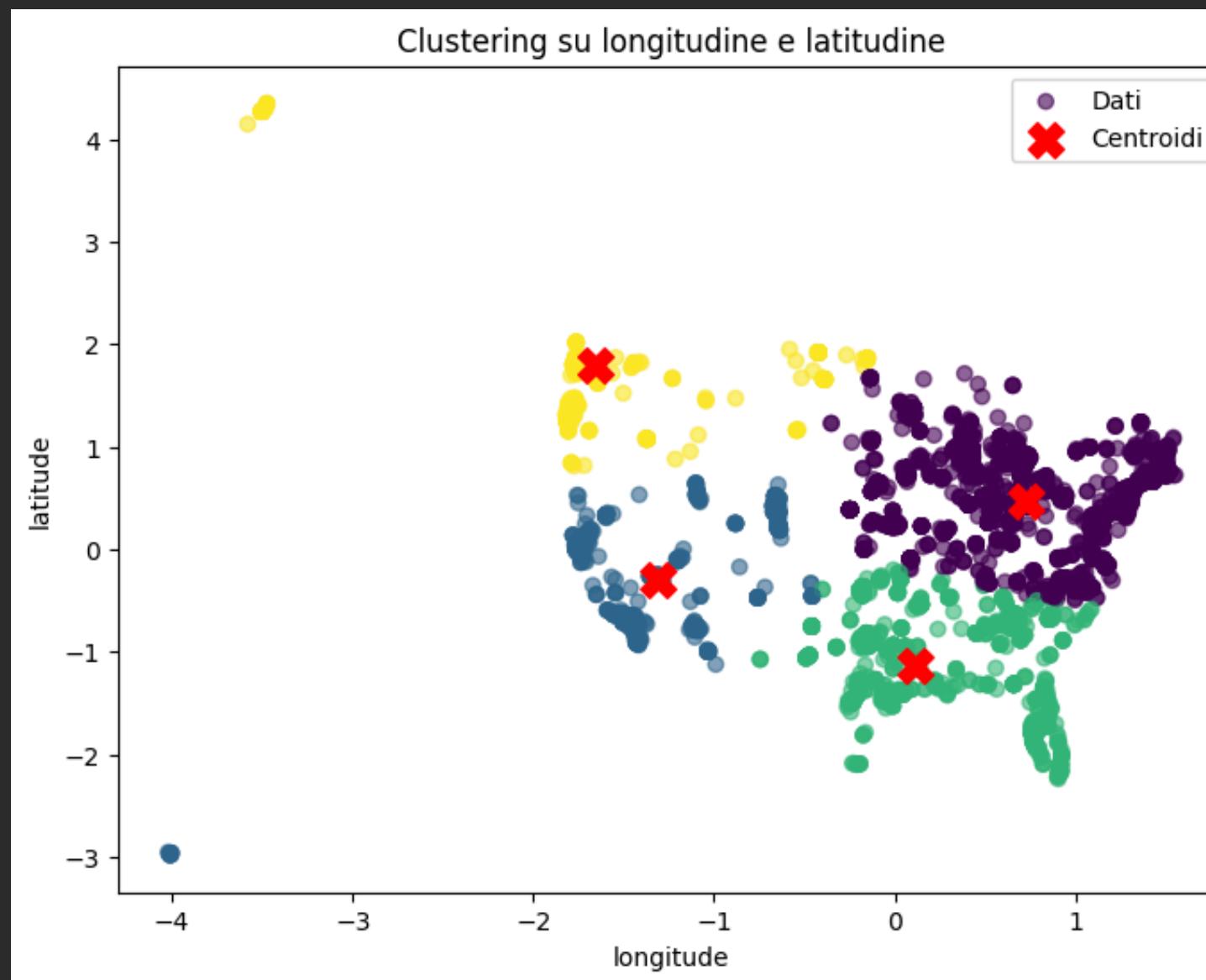
L'analisi è sviluppata sugli stessi parametri utilizzati per il K-Means. Anche in questo caso i dati sono stati normalizzati, portando ad una distorsione del territorio terrestre. DB-Scan genera più cluster, questo mostra come possano esserci più tipi di case con prezzo e metri quadri simili.

Permette di identificare città o quartieri con una struttura di prezzi specifica e separare zone con elevata concentrazione di affitti.

Il grafico di output permette di ottenere informazioni più specifiche ma richiede più tempo per essere analizzato

Clustering geografico

K-Means



Clustering geografico

K-Means

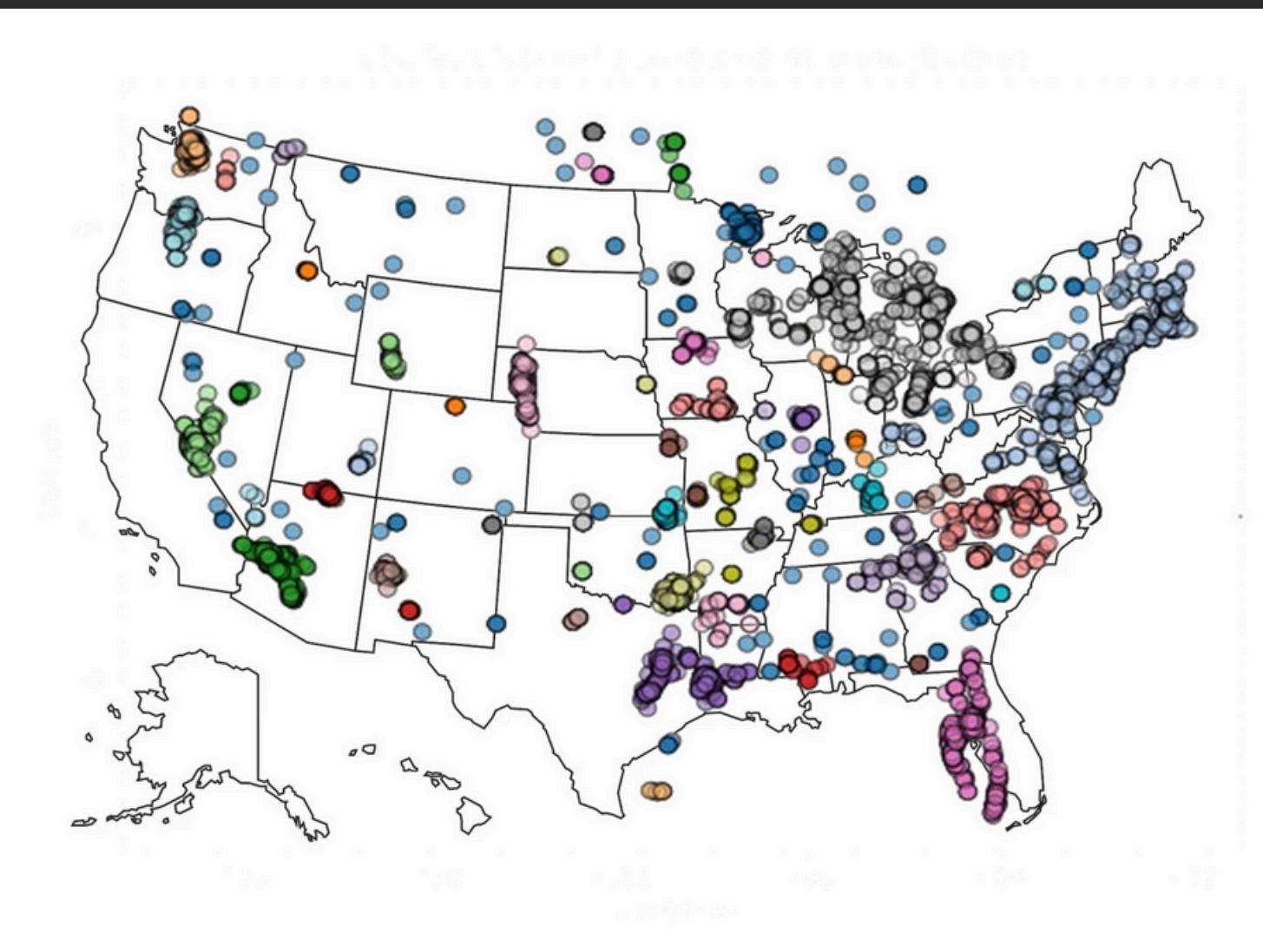
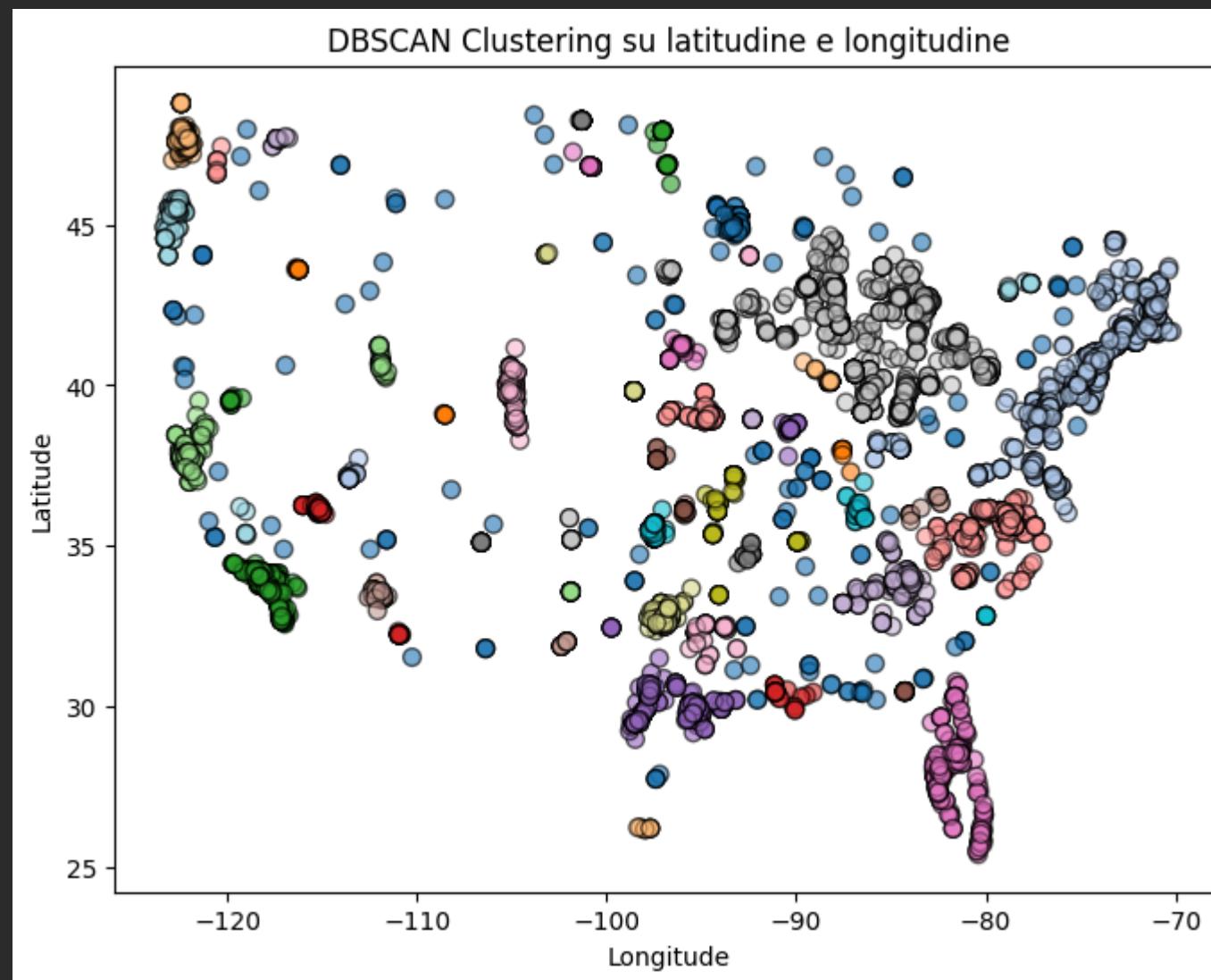
Si ottiene la distribuzione delle case divise in sotto-aree geografiche degli Stati Uniti, rispettivamente:

- nord-est
- nord-ovest
- sud-est
- sud-ovest

Rappresenta una suddivisione abbastanza simmetrica nella sua forma e quindi semplice da intendere. La problematica di un grafico di questo tipo potrebbe essere legata all'assenza di possibilità di individuare densità urbana specifica

Clustering geografico

DB-Scan



Clustering geografico

DB-Scan

Si ottengono molti più cluster, permettendo di poter eseguire un'analisi più specifica anche dal punto di vista della densità. Nel K-Means tutta la east-coast sembra altamente densa, mentre da qui si riescono ad intravedere aree, come l'Alabama e il Mississipi, decisamente meno dense rispetto alla media.

In generale, però entrambi non sono riusciti ad individuare le zone con mancanza di offerta, come la maggior parte degli stati centrali