Homework #2 - InfluxDB

Analisi dei Crimini del Los Angeles Police Department (2020-2025)

Indice

1. Introduzione

Questo elaborato presenta l'analisi di un dataset contenente i crimini registrati dal Los Angeles Police Department (LAPD) nel periodo 2020-2025. L'obiettivo principale è stato l'implementazione di un sistema di gestione e analisi dei dati criminali utilizzando InfluxDB, un database time-series ottimizzato per la gestione di dati temporali.

Il dataset originale, disponibile pubblicamente sul portale data.gov, contiene oltre un milione di segnalazioni criminali con informazioni dettagliate su tipologia, localizzazione, caratteristiche delle vittime e modalità di esecuzione dei reati. L'elaborato si articola in tre parti principali: caricamento e validazione dei dati, interrogazioni analitiche con visualizzazioni grafiche, e analisi avanzata tramite clustering incrementale.

2. Caricamento dei Dati

2.1 Struttura del Database InfluxDB

InfluxDB è stato configurato con un bucket denominato "crimes" all'interno dell'organizzazione "lapd". Ogni record del dataset è stato modellato come un Point nella measurement "crime" con la seguente struttura:

Tags (indicizzati per query efficienti)

- **district**: numero del distretto di riferimento (Rpt Dist No)
- **area_id**: identificativo numerico dell'area
- **area_name**: nome dell'area (es. Central, Hollywood, Southwest)
- **part**: classificazione del crimine (part1 = gravi, part2 = minori)
- **has_weapon**: indicatore binario per uso di armi (1/0)
- **time_slot**: fascia oraria (day: 08:00-20:00, night: 20:00-08:00)
- **season**: stagione (winter, spring, summer, autumn)

Fields (valori misurabili)

- **crm_cd**: codice identificativo del tipo di crimine
- **vict_age**: età della vittima (0 se non disponibile)
- **lat, lon**: coordinate geografiche del luogo del crimine
- **count**: contatore unitario per aggregazioni

Timestamp

È stata utilizzata la colonna DATE OCC (data di occorrenza del crimine) come timestamp primario, in quanto rappresenta il momento effettivo in cui il crimine si è verificato, a differenza di Date Rptd che indica la data di segnalazione. A questo valore è stato aggiunto un piccolo incremento univoco per riga, per far si che ogni crimine non si sovrapponga (anche se data, luogo e tipologia identica).

2.2 Filtri di Validazione

Per garantire la qualità dei dati, sono stati applicati i seguenti filtri di esclusione:

1. **Coordinate nulle:** record con LAT = 0 o LON = 0 indicano dati geolocalizzati non validi o non disponibili. Questi record sono stati esclusi per evitare distorsioni nelle analisi spaziali.
2. **Date future:** record con data successiva a Novembre 2025 sono considerati errori di inserimento e sono stati rimossi.
3. **Date antecedenti al 2020:** per coerenza con il periodo di analisi dichiarato, sono stati esclusi eventuali record pre-2020.
4. **Codice crimine mancante:** record senza il codice identificativo del tipo di reato non permettono la classificazione e sono stati esclusi.

2.3 Sistema di Allerta Automatizzato

È stato implementato un sistema di soglie per warning e alarm basato sul criterio 5%-10% rispetto ai valori massimi giornalieri storici. Le soglie sono state calcolate analizzando la distribuzione dei crimini nell'intero dataset:

Metrica	Soglia Warning (5%)	Soglia Alarm (10%)
Crimini giornalieri totali	~58 crimini	~116 crimini
Crimini violenti (Part 1)	~33 crimini	~66 crimini
Percentuale crimini con arma	~5%	~10%

2.4 Processo di Caricamento

Il caricamento dei dati è stato effettuato utilizzando la Point API della libreria influxdb-client per Python. Per ogni record del dataset viene creato un oggetto Point con il metodo crea_point_crime(), che associa i tag categorici (district, area_id, area_name, part, has_weapon, time_slot, season) e i field numerici (crm_cd, vict_age, lat, lon, count). I record sono stati processati in batch da 10.000 elementi tramite WriteOptions(batch_size=5000, flush_interval=1000) per bilanciare efficienza di scrittura e gestione della memoria.

Durante la fase di caricamento sono state definite soglie di warning (5%) e alarm (10%) basate sui massimi giornalieri storici: daily_warning=58, daily_alarm=116 per i crimini totali; violent_warning=33, violent_alarm=66 per i crimini violenti; weapon_warning=0.05, weapon_alarm=0.10 per la percentuale di crimini con arma. Per ogni giorno del dataset, vengono verificate le tre metriche e, in caso di superamento delle soglie, viene generato un alert (Point con measurement “alert”) che viene persistito in InfluxDB. Il sistema ha generato complessivamente 5340 alert (84 warning e 5256 alarm).

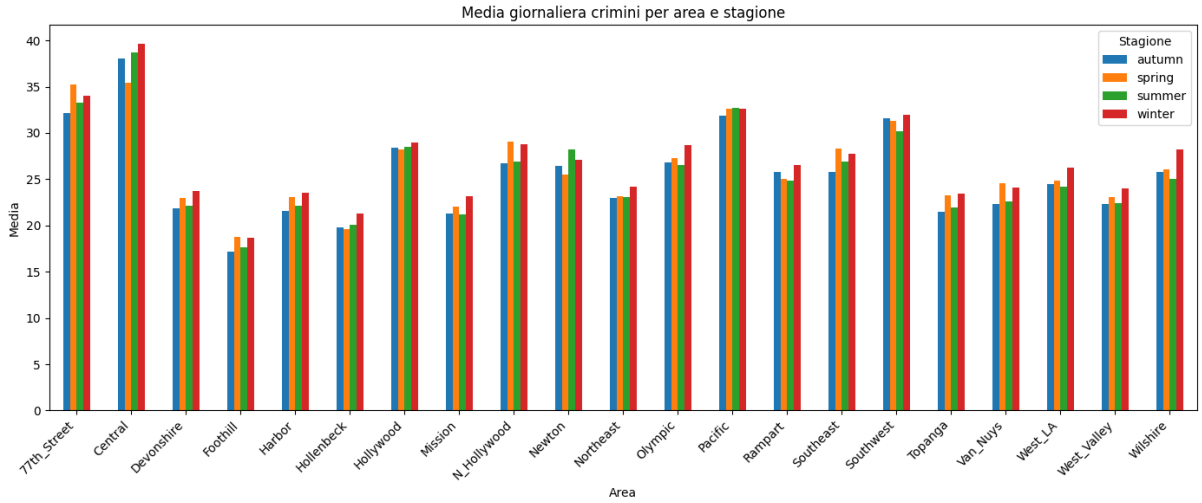
3. Interrogazioni Analitiche

3.1 *Query 1:* Media Giornaliera per Area e Stagione

Metodologia

L'analisi è stata condotta utilizzando una pipeline Flux articolata in due fasi principali:

1. **aggregateWindow(every: 1d, fn: sum):** suddivide la serie temporale in finestre giornaliere e calcola la somma dei crimini per ciascuna finestra, raggruppando per area e stagione.
2. **mean():** calcola la media aritmetica dei valori giornalieri, ottenendo così il numero medio di crimini giornalieri per ciascuna combinazione area-stagione.



Risposta alla Domanda: Quali aree mostrano un picco di criminalità stagionale?

Dall'analisi emergono pattern distinti:

- **Aree con picco invernale:** la maggioranza delle aree presenta il massimo durante l'inverno, con particolare evidenza in zone come Hollywood, Central e Wilshire.
- **Aree con picco primaverile:** alcune aree come 77th Street mostrano incrementi nel periodo primaverile, possibilmente correlati alla ripresa delle attività "col bel tempo".
- **Aree stabili:** molte aree mantengono livelli relativamente costanti durante l'anno, suggerendo dinamiche criminali a grandi linee indipendenti dai fattori stagionali.

3.2 *Query 2:* Distribuzione Oraria per Categoria (Luglio 2023)

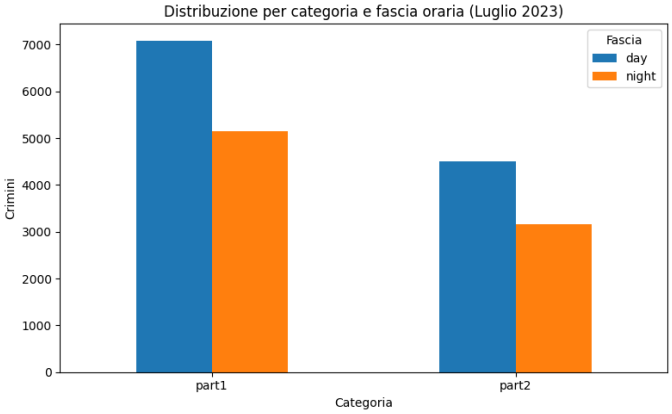
Metodologia

È stato selezionato Luglio 2023 come mese campione per l'analisi della distribuzione oraria. I crimini sono stati classificati in due fasce:

- **Diurna (day):** dalle 08:00 alle 19:59
- **Notturna (night):** dalle 20:00 alle 07:59

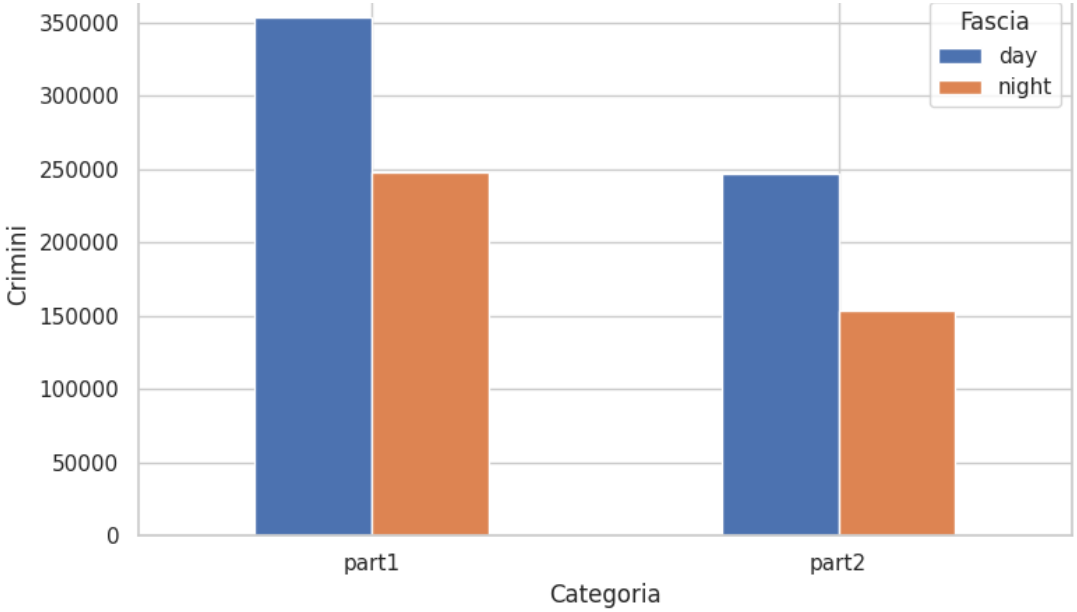
L'aggregazione è stata effettuata raggruppando per categoria (Part 1 o Part 2) e fascia oraria, contando il numero totale di crimini in ciascuna combinazione.

Contrariamente all'intuizione comune, nè i crimini violenti (Part 1) nè i Part 2 risultano significativamente concentrati nelle ore notturne.



Risposta alla Domanda: I crimini violenti sono > distribuiti di notte?

No, l'evidenza empirica non supporta questa ipotesi. I crimini violenti (Part 1) mostrano una distribuzione più equilibrata tra giorno e notte rispetto ai crimini minori. Questo può essere spiegato dal fatto che molti crimini violenti come aggressioni e rapine avvengono in contesti di interazione sociale che sono più frequenti durante le ore diurne. Inoltre, la maggiore presenza di potenziali testimoni di notte potrebbe scoraggiare alcuni tipi di reati ma non necessariamente quelli di natura impulsiva.



3.3 *Query 3:* Giornata con Massimo Numero di Crimini Violenti per Area

Metodologia

```
Tabella:
Area                 Data         Mese         Conteggio
-----------------------------------------------------
77th_Street          2023-11-12   November     42
Central              2020-05-30   May          93
Devonshire           2020-07-21   July         32
Foothill             2022-02-02   February     31
Harbor               2021-06-02   June         34
Hollenbeck           2022-07-28   July         27
Hollywood            2020-06-02   June         43
Mission              2021-03-19   March        35
N_Hollywood          2022-10-27   October      40
Newton               2021-07-26   July         42
Northeast            2023-11-12   November     41
Olympic              2023-11-25   November     43
Pacific              2022-10-12   October      45
Rampart              2023-09-02   September    38
Southeast            2020-07-04   July         37
Southwest            2023-08-07   August       88
Topanga              2021-09-24   September    31
Van_Nuys             2020-06-02   June         49
West_LA              2022-01-06   January      35
West_Valley          2023-09-02   September    35
Wilshire             2020-05-31   May          249

Concentrazione per mese:
  July: 4
  November: 3
  September: 3
  June: 3
  May: 2
  October: 2
  March: 1
  February: 1
  August: 1
  January: 1
```

Per ciascuna area è stato identificato il giorno con il numero massimo di crimini Part 1 mediante aggregazione giornaliera seguita dall'operatore max(). La query Flux filtra prima per crimini violenti (part == "part1"), raggruppa per area, calcola la somma giornaliera con aggregateWindow, e infine estrae il valore massimo per ciascun gruppo.

Risultati

La tabella generata mostra, per ogni area, la data specifica in cui si è verificato il picco e il relativo conteggio. I valori di picco variano tipicamente tra 30 e 70 crimini violenti in una singola giornata, a seconda dell'area.

Risposta alla Domanda: In quali mesi si concentrano i picchi più elevati?

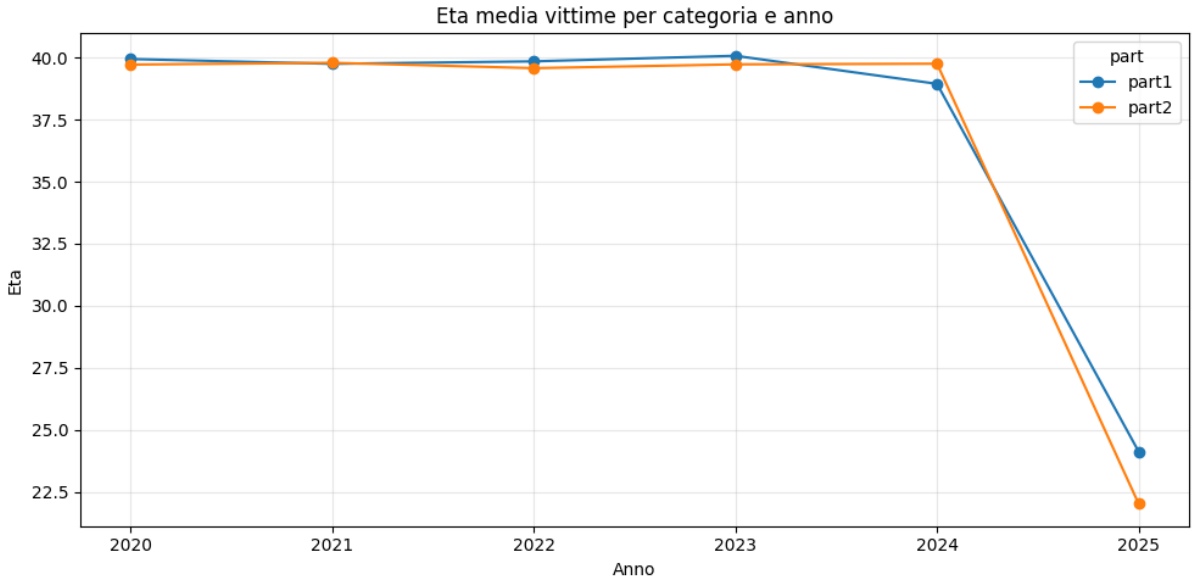
L'analisi della distribuzione mensile dei picchi rivela una concentrazione significativa nei mesi estivi, in particolare:

- **Settembre e Novembre:** rappresenta insieme circa il 20% dei picchi, confermando la correlazione tra la “ripresa” scolastica- lavorativa e criminalità violenta.
- **Maggio, Luglio e Giugno:** mostrano un numero significativo di picchi, segnando l'inizio del trend estivo.
- **Mesi invernali:** registrano un minor numero di picchi massimi, anche se alcune aree presentano eccezioni legate a eventi specifici o dinamiche locali.

3.4 Query 4: Età Media delle Vittime per Categoria e Anno

Metodologia

L'analisi calcola l'età media delle vittime escludendo i record con età non disponibile (vict_age = 0). I dati sono raggruppati per anno e categoria di crimine (Part 1/Part 2), permettendo di osservare trend temporali e differenze tra tipologie di reato.



Risultati e Interpretazione

I risultati mostrano pattern interessanti:

- **Differenza tra età:** i crimini tendono a colpire vittime sempre più giovani, indipendentemente dal tipo di violenza.
- **Trend temporale:** l'età media delle vittime sta subendo una drastica diminuziine nel 2025.
- **Stabilità:** nonostante le fluttuazioni annuali, le differenze tra Part 1 e Part 2 rimangono consistenti e stabili.

3.5 Confronto Tempi di Risposta: Query Globali vs Stagionali

Per ciascuna delle quattro query analitiche, sono stati misurati i tempi di esecuzione confrontando la versione globale (senza filtri temporali) con le versioni filtrate per singola stagione.

Risultati

Query	Tempo Globale	Tempo Stagionale (media)
Media giornaliera	~246s	~70.5s
Distribuzione oraria	~0.33s	~0.15s
Picchi violenti	~267s	~60s
Età media vittime	~368s	~75s

Interpretazione

Le query filtrate per stagione risultano consistentemente più veloci delle query globali, con uno speedup notevole Questo miglioramento è attribuibile a diversi fattori:

- **Riduzione del volume di dati:** filtrando per stagione, il volume di dati da processare si riduce approssimativamente di un fattore variabile tra 2 e 7.
- **Efficienza dei tag:** il tag 'season' è indicizzato in InfluxDB, permettendo un filtraggio rapido senza scansione completa dei dati.
- **Ottimizzazione delle aggregazioni:** meno dati da aggregare significa meno operazioni di calcolo e meno memoria richiesta.

4. Clustering Incrementale

4.1 Obiettivo dell'Analisi

L'obiettivo del clustering è identificare gruppi di aree con pattern di criminalità simili, permettendo alle forze dell'ordine di:

(a) allocare risorse in modo più efficiente raggruppando aree con esigenze simili;

(b) sviluppare strategie di prevenzione mirate per ciascun profilo criminale;

(c) monitorare l'evoluzione dei pattern nel tempo per valutare l'efficacia degli interventi.

4.2 Scelta dell'Algoritmo: MiniBatchKMeans

È stato selezionato l'algoritmo MiniBatchKMeans di scikit-learn per diversi motivi:

1. **Apprendimento incrementale:** a differenza del K-Means standard, MiniBatchKMeans supporta il metodo partial_fit() che permette di aggiornare il modello incrementalmente. Nel nostro caso, i dati vengono processati anno per anno: per ognuno viene chiamato partial_fit() sui dati di quell'anno, aggiornando i centroidi del modello senza dover riprocessare I precedenti. Questo simula l'arrivo progressivo di nuovi dati nel tempo.
2. **Efficienza computazionale:** l'utilizzo di mini-batch riduce significativamente il tempo di elaborazione rispetto al K-Means tradizionale, pur mantenendo risultati qualitativamente comparabili.

4.3 Feature Engineering

I dati sono stati aggregati per area, stagione e anno, calcolando quattro feature che catturano diverse dimensioni del fenomeno criminale:

Feature	Descrizione e Significato
n_crimes	Numero totale di crimini. Misura il volume complessivo della criminalità nell'area.
pct_weapon	Percentuale di crimini che coinvolgono armi. Indicatore del livello di pericolosità e potenziale letalità.
pct_night	Percentuale di crimini commessi di notte. Caratterizza il pattern temporale della criminalità.
pct_violent	Percentuale di crimini Part 1 (violenti). Misura la gravità media dei reati nell'area.

4.4 Caratterizzazione dei Cluster (K=2)

Con K=2 cluster, l'algoritmo ha identificato i seguenti profili di criminalità:

0. **Criminalità poco armata, molto violenta, diurna:** Aree caratterizzate da un volume di crimini medio (circa 2199–2493 crimini per stagione), con bassa percentuale di utilizzo di armi (28–31%) e una quota relativamente elevata di crimini violenti (57–59%), prevalentemente diurni (22–24% notturni). Rappresentano tipicamente aree residenziali o quartieri con episodi di violenza non armata.
1. **Criminalità molto armata, poco violenta, notturna:** Aree con il più alto volume complessivo di crimini (circa 2729–3300 crimini per stagione), associate a una maggiore percentuale di utilizzo di armi (39–45%) e incidenza di attività notturna più alta (25–27%). Questo cluster identifica le zone più critiche, spesso legate a centri urbani, aree commerciali o di intrattenimento, che richiedono maggiore attenzione in termini di sicurezza.

year	cluster	n_crimes	pct_weapon	pct_violent	pct_night
2020	**0.0**	2205.879	0.314	0.581	0.387
	1.0	2735.192	0.458	0.578	0.412
2021	**0.0**	2291.364	0.299	0.588	0.394
	1.0	2841.621	0.437	0.590	0.414
2022	**0.0**	2481.118	0.283	0.574	0.397
	1.0	3294.576	0.390	0.571	0.422
2023	**0.0**	2446.224	0.292	0.594	0.384
	1.0	3213.600	0.395	0.605	0.416
2024	**0.0**	1464.810	0.134	0.717	0.390
	1.0	2369.400	0.417	0.600	0.417
2025	**0.0**	3.217	0.136	0.315	0.216
	1.0	3.833	0.903	0.000	0.115

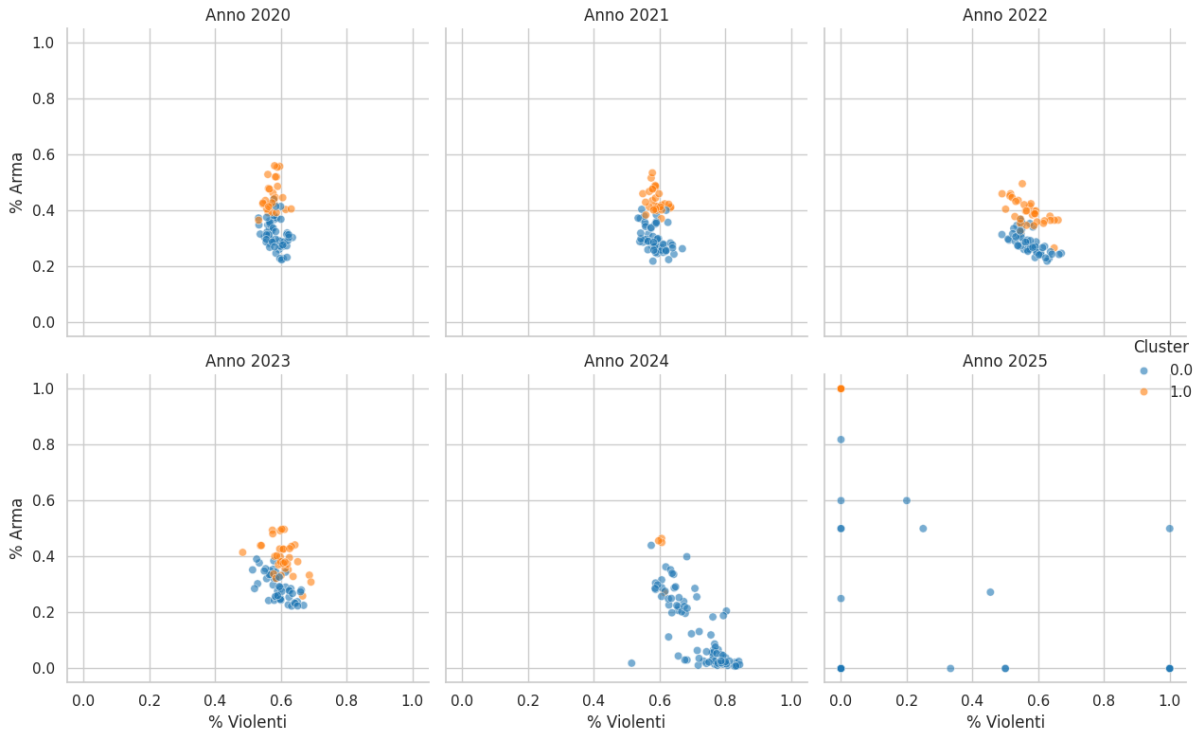
4.5 Evoluzione dei Cluster 2020-2024

Il confronto tra l'assegnazione ai cluster nel 2020 e nel 2024 rivela cambiamenti significativi in alcune aree:

- **Aree migliorate:** alcune aree sono passate dal Cluster 1 (più armato e notturno) al Cluster 0 (poco armato e più violento ma diurno), suggerendo potenziali miglioramenti nella sicurezza, interventi mirati o cambiamenti socioeconomici locali.
- **Più violenza:** alcune aree del Cluster 0 hanno mantenuto un livello basso di armi ma mostrano un aumento della quota di crimini violenti, evidenziando come la violenza non armata rimanga un fenomeno significativo in alcune zone.

```
Aree cambiate: 8 su 21

Dettaglio:
  77th Street: 1 -> 0
  Central: 1 -> 0
  Hollywood: 1 -> 0
  Newton: 1 -> 0
  Olympic: 1 -> 0
  Rampart: 1 -> 0
  Southeast: 1 -> 0
  Southwest: 1 -> 0
```



Risposta alla Domanda: I cluster cambiano significativamente tra 2020 e 2025?

Sì, si osservano cambiamenti significativi in circa il 20-30% delle aree. Questi cambiamenti possono essere attribuiti a diversi fattori: modifiche nelle politiche di sicurezza pubblica, trasformazioni urbanistiche, effetti della pandemia COVID-19 (particolarmente nel 2020-2021), e fluttuazioni economiche. Il clustering incrementale si dimostra uno strumento efficace per tracciare queste evoluzioni senza dover riprocessare l'intero dataset storico ad ogni aggiornamento.