**Operazioni Effettuate ‘SPA\_2024\_per\_OpenDATA.csv’**

Lo scopo di questo documento è descrivere in modo dettagliato il processo di pre-processing applicato al dataset ‘SPA\_2024\_per\_OpenDATA.csv’, relativo alle Prestazioni Specialistiche Ambulatoriali per l’anno 2024.  
Il trattamento dati è volto a garantire qualità, coerenza, chiarezza e facilità d’uso per la pubblicazione come Open Data, favorendo la fruibilità da parte di utenti e stakeholder.

**--- Inizio Pre-processing ---**

**Numero di righe: 11149**

**Numero di colonne: 4**

**Dimensioni del dataset (righe, colonne): (11149, 4)**

**Max Nr Prestazioni : 181637 , Min: 1**

**--- Fine Pre-processing ---**

**Numero di righe: 9927**

**Numero di colonne: 6**

**Dimensioni del dataset (righe, colonne): (9927, 6)**

**Max Numero Prestazioni Erogate : 181637 , Min: 1**

**1. Ridenominazione delle Colonne**

La prima fase ha previsto la rinomina delle colonne del dataset con l’obiettivo di uniformare la nomenclatura e facilitare la comprensione dei dati.  
È stata applicata una mappatura delle intestazioni originali verso una forma standardizzata:

**' NEW CODICE CATALOGO -> ‘Nuovo Codice Catalogo',**

**' NEW DESCR CATALOGO -> 'Nuova Descrizione Catalogo',**

**'Accesso'->'Accesso'**

**'Nr Prestazioni’ -> 'Numero Prestazioni Erogate'**

Questa operazione assicura che tutte le colonne abbiano nomi coerenti e privi di ambiguità, facilitando operazioni successive come la manipolazione e l’analisi.

**2. Rimozione di Spazi Superflui**

È stata effettuata un’accurata pulizia dei dati testuali presenti nel dataset. Questo intervento ha previsto:

* Rimozione di spazi bianchi iniziali e finali
* Eliminazione di caratteri speciali indesiderati che potrebbero compromettere le analisi o la visualizzazione

Dall’analisi delle colonne testuali è emerso che nessun valore presentava spazi superflui o anomalie. Sebbene non siano state necessarie modifiche sostanziali, questa fase ha consolidato la qualità complessiva del dataset, prevenendo possibili errori invisibili che avrebbero potuto influire sulle elaborazioni future.

**3. Uniformazione del Case (Title Case)**

Per migliorare la coerenza visiva e la leggibilità dei dati testuali, è stata applicata l’uniformazione dei valori nel formato Title Case (prima lettera maiuscola di ogni parola).

Questa scelta:

* Riduce le discrepanze causate da formattazioni incoerenti
* Favorisce una migliore esperienza d’uso in consultazioni e report
* Facilita l’identificazione univoca dei valori categoriali

L'uniformazione del case è stata applicata a 4 colonne del dataset. Pur essendo un intervento che non ha comportato un elevato numero di modifiche puntuali, contribuisce significativamente a migliorare la qualità semantica e l’ordine del dataset.

**4. Rimozione di Righe con Specifici Valori**

È stata implementata una fase specifica per la rimozione delle righe in cui la colonna ‘Nuova Descrizione Prestazione’ conteneva un punto ('.'). Questa operazione è stata eseguita per eliminare potenziali descrizioni incomplete o non conformi.

* Numero di righe rimosse: 1222
* Numero di righe nel DataFrame pulito dopo questa operazione: 9927

Questa fase mirata ha contribuito a raffinare ulteriormente la qualità del dataset, eliminando voci potenzialmente problematiche per l'analisi.

**5. Rimozione di Duplicati**

È stata effettuata una verifica approfondita per individuare eventuali righe duplicate, al fine di evitare distorsioni nelle analisi statistiche, aggregazioni o modelli predittivi.

La scansione ha esaminato l’intero dataset, confrontando ogni riga per rilevare duplicazioni esatte.

Risultato: nessuna riga duplicata è stata trovata, confermando l’unicità di ogni osservazione e garantendo così l’integrità del dataset.

Questa verifica, pur senza richiedere azioni correttive, è un passaggio chiave per assicurare la qualità e l’affidabilità complessive dei dati.

**6. Gestione di Valori Nulli**

È stata effettuata una scansione completa delle colonne principali per individuare valori nulli o mancanti.  
Il risultato ha mostrato che tutte le colonne erano complete e non contenevano dati assenti.  
Questo rappresenta un ottimo punto di partenza per l’analisi, in quanto non sono state necessarie imputazioni o rimozioni.

**7. Rimozione Righe con Descrizioni Anomale**

È stata applicata una logica di filtro per eliminare le righe la cui colonna Nuova Descrizione Catalogo contenesse il carattere punto (.), considerato in questo contesto un indicatore di inconsistenza semantica o formattazione errata.  
Sono state rimosse 1222 righe, portando il totale delle osservazioni da 11149 a 9927 righe.

**Arricchimento del Dataset**

**8. Creazione della Colonna "Tipo Prestazione"**

Sono state introdotte due nuove variabili derivate per arricchire il dataset:

* **Descrizione Completa Prestazione:** Questa colonna è stata creata concatenando il contenuto delle prime tre colonne ('Nuovo Codice Prestazione Catalogo', 'Nuova Descrizione Prestazione', 'Accesso'). L'obiettivo è fornire una descrizione testuale più completa per ogni prestazione.
* **Frequenza Erogazione:** Questa variabile è stata derivata dalla colonna ‘Numero Prestazioni Erogate’ mediante un processo di binning. La logica adottata prevede la suddivisione dei valori numerici in fasce di frequenza, definite sulla base della distribuzione dei dati:
  + Bassa: da 1 a 2 prestazioni erogate
  + Media: da 3 a 107 prestazioni erogate
  + Alta: da 108 a 181637 prestazioni erogate

Obiettivi e Benefici:

* + Riduzione della complessità numerica nella variabile originaria
  + Facilitazione nell’analisi comparativa tra gruppi omogenei di prestazioni
  + Introduzione di una dimensione qualitativa utile per l’esplorazione dei dati e la modellazione predittiva

L’arricchimento si è concluso con successo, aggiungendo le colonne ‘Descrizione Completa Prestazione’ e ‘Frequenza Erogazione’ al dataset.

**9. Esportazione Finale**

Al termine delle operazioni di pulizia, trasformazione e arricchimento, il dataset è stato esportato nel formato CSV con il nome: ‘Prestazioni Specialistiche Ambulatoriali 2024.csv’.

Motivazioni della scelta del formato CSV:

* Massima compatibilità con software di analisi dati, fogli di calcolo (Excel, Google Sheets), strumenti di visualizzazione e ambienti di programmazione (Python, R, etc.)
* File leggero e facilmente leggibile anche da utenti non esperti
* Formato standard aperto, ideale per la pubblicazione come Open Data, senza dipendenza da software proprietari

Questa operazione rappresenta un passaggio cruciale nella pipeline di gestione dati, in quanto:

* Consolida tutte le modifiche effettuate in una versione definitiva e documentata.
* Rende il dataset pronto per la distribuzione, l’analisi o l’integrazione in portali open data e sistemi informativi.
* Garantisce trasparenza e tracciabilità dei dati trattati, supportando la riproducibilità delle analisi future.

Il dataset così processato è stato esportato in un file finale, pronto per la pubblicazione come Open Data. Il formato è stato scelto per garantire ampia compatibilità e facilità di accesso, accompagnato da documentazione esplicativa per agevolare la comprensione da parte degli utenti.

**Conclusioni**

Il pre-processing del dataset ha garantito una pulizia completa dei dati e l’arricchimento delle informazioni tramite nuove colonne derivate.  
Le trasformazioni effettuate aumentano la qualità, leggibilità e usabilità del dataset e pongono una solida base per l’analisi esplorativa e modellistica.  
Ogni passaggio è stato tracciato e validato per assicurare l’affidabilità e la ripetibilità del processo.