Esercizio 1

Analisi delle anomalie più rilevanti nel dataset:

1. Errori di formato nelle colonne e valori NaN

I formati dei placeholder presentano numerose criticità. Di seguito quelle che maggiormente incidono sulla qualità dei dati.

1.1

La colonna *data* risulta essere la più problematica, con un totale di 619 anomalie rilevate in fase di analisi, tra cui 4 valori NaN.

```
"Totale anomalie: 619 anomalie
Anomalie separate:
valori_nani: 4
anomalie date_non_conformi: 615 anomalie"
```

Su 145 formati corretti, i cluster con più alta incidenza sono:
- *"00-00-0000"* con 283 ricorrenze;
- *"00/00/00"* con 204.

1.2

Anche la colonna *totale* presenta molteplici anomalie: 480 complessivamente, di cui 3 valori NaN.

```
"Totale anomalie: 480 anomalie
Anomalie separate:
valori_nani:3
anomalie totale_non_conformi: 477 anomalie"
```

Su 284 formati corretti, le anomalie di formato più frequenti sono:
- *"0,00"* che appare 176 volte;
- *"00.00"*, 68.

1.3
Segue la colonna *ora* che presenta 155 anomalie e 9 valori NaN.

```
"Totale anomalie: 155 anomalie
Anomalie separate:
valori_nani: 9
anomalie ora_non_conformi: 146 anomalie"
```

Gli errori di formato più ricorrenti sono:
- *"00:00:00"* che appare 86 volte
- *"00.00.00",* 11.

1.4

Benché i NaN abbiano un'incidenza bassa, sono presenti in tutte le colonne del dataset (eccetto in filename). La colonna *ora* ha l'incidenza maggiore con 9 valori NaN.

Seguono:
- *data* con 4;
- *totale* con 3;
- *partita iva* con 2.

2. Piano di validazione automatica

2.1

Utilizzare un DataFrame che consenta di applicare espressioni regolari per rilevare la sistematicità di anomalie nei formati e, per lo più, gestirle automaticamente.

i) Volendo mantenere i placeholder:

- Data: ^00/00/0000$

- Ora: ^00:00$

- Totale: ^00,00$

- Partita iva: ^00000000000$

ii) Assumendo che i placeholder siano già stati sostituiti con valori reali:

- Data: ^\d{2}/\d{2}/\d{4}$

- Ora: ^\d{2}:\d{2}$

- Totale: ^\d{2},\d{2}$

- Partita iva: ^\d{11}$

iii) Oppure combinandole:

- ^(?!00/00/0000$)\d{2}/\d{2}/\d{4}$

- ^(?!00:00$)\d{2}:\d{2}$

- ^(?!00,00$)\d{2},\d{2}$

- ^(?!00000000000$)\d{11}$

2.2

Per determinare il margine d'errore tollerabile nel dataset, sarebbe necessario analizzarne le variabili e stabilirne la priorità date dal contesto specifico. Tuttavia, essendo il dataset composto da un numero relativamente ridotto di dati, anche poche anomalie potrebbero comprometterne la qualità complessiva. Pertanto proporrei di impostare delle soglie di tolleranza basse, circa all'0,5% o 1%.

3. Report periodico di validazione

3.1

i) Descrivere contesto e tipo di dati da analizzare;

ii) Delineare gli obiettivi;

iii) Definire il periodo di riferimento del report.

3.2

i) Sintesi dei controlli eseguiti: descrivere dettagliatamente il tipo di controlli effettuati sul campione dei dati (che siano codici o controlli manuali);

ii) Riepilogo delle anomalie riscontrate per ciascuna colonna e gravità delle anomalie;

iii) Qualora fosse possibile, determinare la causa delle anomalie: analizzarene i pattern o le modifiche che potrebbero avere generato errori (nel nostro caso, l'inserimento manuale dei dati).

3.3

i) Monitoraggio diacronico delle variazioni: creare un DataFrame per il controllo periodico di eventuali nuove anomalie, che ne riporti la percentuale di incidenza per ciascuna colonna - data; ora; totale e partita iva - e che permetta di fare un confronto con gli altri periodi;

ii) Un grafico per valutare le variazioni nel tempo.

3.4

i) Attuare azioni correttive attraverso un piano di validazione automatico.

3.5

i) Conclusioni e future azioni da intraprendere.