Struttura dataset

1. Introduzione

Questo documento ha lo scopo di descrivere la struttura del dataset che andrò ad analizzare in maniera da riuscire a comprendere meglio i passaggi e definire che rulo hanno i dati.

In questo caso i dati non sono raccolti attraverso database o un data lake, ma per problemi logistici abbiamo dovuto utilizzare un dataset in formato csv reperito online, che, riporta ordini degli utenti di un e-commerce online.

2. Struttura

I dati sono strutturati in maniera tabellare in quanto in formato .csv e poichè non presentano dati come immagini, audio ecc…

```
df.head(10)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047.0	United Kingdom

Figure 1: Struttura del dataset

La struttura si presenta nel seguente modo, andremo ad analizzare quindi le colonne e i valori che rappresentano:

- **InvoiceNo**: rappresenta il numero di fattura per ogni acquisto, è possibile quindi che ci siano fatture uguali per articoli differenti. E' di tipo oggetto, questo significa che potrebbe non essere unicamente numerico;
- **StockCode**: rappresenta il numero dell'oggetto associato ad esso, è quindi un numero univoco per oggetto. Anch'esso è di tipo oggetto;
- **Description**: rappresenta la descrizione degli oggetti acquistati, è di tipo oggetto;
- **Quantity**: rappresneta la quantità dell'oggetto acquistato in fattura. Questo è un campo di tipo intero;
- **InvoiceDate**: rappresenta la data e l'ora di acquisto, è di tipo oggetto, facilmente convertibile in tipo date;
- **UnitPrice**: rappresenta il prezzo della singola untià, facilmente recuperabile la spesa totale della fattura moltiplicandola per la quantità. E' un dato rappresentato dal tipo float;
- **CustomerID**: rappresenta l'identificativo univoco dell'utente, anche'esso è di tipo float;
- **Country**: rappresenta la nazione di provenienza del cliente, rappresentato da un tipo oggetto.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
```

Figure 2: Rappresentazione tipo di dato