Dipartimento di Informatica

Laurea Magistrale in
Data Science

Analisi delle Rotte Aeree Europee

Birti Mattia 897092, Longo Gloria 864579

09/02/2023

Indice

1. Introduzione

L'intero progetto si basa sull'idea di realizzare un database il più completo possibile che racchiuda tutte le informazioni riguardanti le rotte aeree europee.
L'obbiettivo del progetto consiste nella costruzione e implementazione di un ampio dataset al quale sottoporre in seguito delle interrogazioni per estrarre il maggior numero di informazioni sulle rotte aeree europee, sulle compagnie aeree utilizzate e sui vari aeroporti coinvolti.
Per ottenere questa grande quantità di dati sono state utilizzate diverse tecniche tra le quali il web scraping (applicato su due siti web distinti), la web API (per l'estrazione del maggior numero dei dati riguardanti le rotte aeree) e infine un semplice download (file .csv) utile ai fini dell'integrazione del dataset.
È stato in seguito deciso di utilizzare un database di tipo NO-SQL data la natura dei dati a nostra disposizione e visti gli enormi vantaggi che esso mette a disposizione tra i quali la possibilità di andare oltre lo schema relazionale.
Il modello NO-SQL scelto è di tipo Document Based, precisamente MongoDB, in quanto la sua struttura ad albero permette di maneggiare nel miglior modo i dati a nostra disposizione, inoltre grazie a questa sua caratteristica, l'inserimento e la rimozione di interi documenti o collezioni risultano essere molto semplici (caratteristica molto utile per un possibile aggiornamento futuro su rotte mondiali e non solo europee).
I dati, una volta integrati, sono stati inseriti nel database tramite l'utilizzo della libreria *pymongo* di Python, con la quale si sono svolte tutte le successive operazioni di interrogazione del database.
In questo report verranno dunque illustrati tutti i passaggi attutati per lo sviluppo del progetto, partendo dall'estrapolazione dei dati fino ad arrivare alle conclusioni di natura descrittiva.

2. Sorgenti Dati

Per l'ottenimento dei dati da internet esistono due metodologie differenti: il Web Scraping e l'utilizzo di Web API.

Le API (Application Programming Interface) sono dei protocolli utilizzati come un'interfaccia (che possono esistere sia a livello di rete, sia a livello di dispositivo) che mettono in comunicazione diversi dispositivi, in questo caso mettono in comunicazione due sorgenti software differenti. Esse sono costituite da un insieme di routine, strutture dati o variabili che permettono al programmatore di richiamare le funzionalità di un'applicazione di terze parti. Il formato con cui le API generalmente rispondono è di tipo JSON (JavaScript Object Notation) il quale è tipicamente basato su due strutture: una collezione di coppie nome_valore (chiamata object) e una lista ordinata di valori (array).

Lo Scraping si tratta di una tecnica per il download di informazioni dalle pagine web molto basilare semplicemente navigando all'interno del codice HTML, generalmente eseguito in modo automatizzato tramite degli script, e copiandi le informazioni interessate.

2.1. Web Scraping - Airline

La tecnica di Web Scraping è stata utilizzata da noi per il download della lista di tutte le compagnie aeree europee[1] e per il download della lista di tutte le compagnie aeree mondiali[2] con i relativi codici IATA. In entrambi i casi lo scraping è stato effettuato tramite esecuzione di un jupyter notebook[3] ed il linguaggio di programmazione Python.

[1] https://airmundo.com/it/blog/elenco-delle-compagnie-aeree-in-europa/
[2] https://www.flightradar24.com/data/airlines/
[3] AirlineEUscraping.ipynb

I principali passi eseguiti dal notebook per completare lo scraping sono stati i seguenti:

1. Ottenimento dell'html della pagina dal sito di Airmundo.
2. Ottenimento di tutti i nomi delle compagnie aeree europee e salvate in una list in python.
3. Richiamato e scaricato il codice html della pagina di FlightRadar24.
4. Salvataggio in una list Python di tutte le compagnie aeree mondiali ed i relativi codici IATA.
5. Confrontato le due liste ed ottenuta una singola lista completa contenete solamente le compagnie aeree europee con i relativi codici IATA.
6. Creazione di un unico .csv (Airline, iataAirline).

2.2. Download Web - Airports

È stato poi eseguito un semplice download di un file excel dal sito ufficiale ACI [4] contenente la lista di tutti gli aeroporti europei ed i relativi codici IATA.

2.3. API - Routes

Per l'ottenimento delle rotte dalla pagine di FlightRadar24 [5] è stata utilizzata la tecnica di Web API, e come nel caso precedente, il tutto è avvenuto tramite esecuzione di un jupyter notebook[6] con linguaggio Python.
Vengono quindi riassunti i principali passaggi:

1. È stato caricato il .csv precedentemente ottenuto[7] tramite tecnica di Web Scraping per avere la lista aggiornata di tutte le compagnie aeree europee e i relativi codici IATA.
2. È stata caricata la lista contenente tutti i codici IATA degli aeroporti europei ed i rispettivi paesi[8].

[4] https://store.aci.aero/product/bundle-2022-edition-traffic-dataset-report/
[5] https://www.flightradar24.com/data/airlines/' + airline + '/routes?get-airport-arr-dep={}
[6] RoutesEUscraping.ipynb
[7] AirlineEU.csv
[8] AirportEU.csv

3. Sono state quindi create diverse variabili di "tempo" per l'ottenimento delle rotte per diverse date temporali (l'intervallo di una settimana esatta).

4. È stato quindi richiamato il sito di FlightRadar24 ed ottenuto in formato JSON le informazioni relativi alle rotte per ogni compagnia aerea e per ogni aereoporto europeo.

5. I risultati così ottenuti, dopo essere stati salvati momentaneamente in diverse liste, sono stati salvati in un file .csv mediante l'uso di pandas contenente più di 13.000 rotte (Airline, Departure, Arrivals, Aircraft, Distance).

3. Esplorazione Dati

La fase di Data Exploration è stata realizzata in un primo momento durante il download dei dati e in un secondo tramite un notebook creato adhoc utilizzando Python con l'ausilio della libreria *pandas* e della libreria *pandas_profiling*.

3.1. Data Cleaning

La fase di Data Cleaning, nel nostro specifico caso, si è svolta in contemporanea con il download primitivo dei dati. Infatti ogni qualvolta i dati venivano scaricati, venivano prima inseriti in una python list e stampati a video in modo da verificare che i dati venissero salvati nel formato e nell'ordine corretto. Nel caso di una riga vuota, oppure di una riga con aereoporti mancanti, veniva scartata a priori. Alla fine di questa fase i dati così ottenuti sono stati poi salvati all'interno di un unico e completo .csv tramite l'utilizzo di Pandas.

3.2. Data Quality

Per la fase di data quality è stato creato un ulteriore notebook[1] con linguaggio python dove sono stati confrontati i risultati ottenuti dal download dei dati relativi alle rotte aeree europee e gli altri due dataset conteneti rispettivamente i paesi europei con i relativi aereoporti e le compagnie aeree.

Si è deciso di verificare e valutare la qualità di tutti e tre i dataset ottenuti tramite la completezza e la ridondanza;

3.2.1 Ridondanza

La prima analisi di qualità effettuata consiste nella verifica dei duplicati a livello di riga. In particolare, per i dataset Airline e Airport non sono stati

[1] DataExploration_Quality.ipynb

rilevati duplicati.
Analizzando invece il dataset ottenuto tramite API (RouteEU.csv), in un primo momento non sono state rilevate righe duplicate, ma con un'analisi più approfondita ed escludendo dall'analisi dei duplicati le colonne relative alla distanza e al modello di aereoplano, sono risultate esserci ben 3025 compagnie aeree con la rotta aerea ridondante. I duplicati sono stati così rimossi ed è stato creato un nuovo dataset con i dati puliti.

3.2.2 Completezza

Relativamente alla caratteristica di completezza, sono state fatte delle analisi a livello di "campo" per ogni tabella.
Per calcolare la completezza è stata utilizzata la seguente formula:

$$Completeness = \frac{NumberMissingValueColumns}{NumberOfRow}$$

Airline Relativamente al csv contenente le compagnie aeree europee ottenuto tramite la tecnica di Web Scraping non risultano esserci dati mancanti.

	Total	Percent
Airline	0	0.0
Airline_Code	0	0.0

Figura 3.1: Missing Value AirlineEU.csv

Airport Stessa cosa per il dataset contenente gli aeroporti europei.

	Total	Percent
Country	0	0.0
Country_Code	0	0.0
City	0	0.0
Airport_Name	0	0.0
Airport_Code	0	0.0

Figura 3.2: Missing Value AirportEU.csv

Route Diverso per il dataset ottenuto tramite API riguardante le rotte aeree europee. Infatti come è stato già detto in precedenza, venivano scartati tutti i risultati che avessero qualche aereopoerto mancante (non avrebbe avuto senso mantenerli) e per questo risulta essere un dataset decisamente completo se si palra dei campi Airline, Departure e Arrivals, ma risulta avere dei dati mancanti sotto la voce Aircraft per un totale di 766 (7.466%).

	Total	Percent
Aircraft	766	0.074666
Airline	0	0.000000
Departure	0	0.000000
Arrivals	0	0.000000
Distance	0	0.000000

Figura 3.3: Missing Value RouteEU.csv

Profiling In conclusione, per avere anche un report relativo alla qualità di tutti e tre i dataset, mediante l'utilizzo della libreria *pandas_profiling*, sono stati creati tre report[2] in formato html che racchiudono ogni tipo di analisi da noi effettuata e non.

[2]AirlineProfiling.html, AirportProfiling.html, RouteProfiling.html

4. Data Integration

In questa sezione del report verrà illustrato il procedimento consistente nell'integrazione del dataset messo in atto con la tecnica di record linkage.
Dapprima il processo è avvenuto con un'operazione di *Corrispondences Investigation*, seguito dalla *Data Preparation* fino ad arrivare al merge tra le diverse sorgenti di dati.

4.1. Descrizione dei Dataset

Il primo dataset che verrà analizzato è quello contenente le compagnie aeree operative in Europa ottenuto con la tecnica del Web Scraping. Esso è composto dai seguenti attributi:

- Airline: nome della compagnia aerea
- Airline_Code: codice univoco della compagnia aerea

Il secondo dataset che verrà descritto è quello contenente le tratte dei voli aerei effettuate nell'ultima settimana di gennaio.
Gli attributi selezionati sempre con la tecnica di Web Scraping sono:

- Airline: nome della compagnia aerea
- Departure: sigla dell'aeroporto di partenza
- Arrivals: sigla dell'aeroporto di arrivo
- Aircraft: modello dell'aereo utilizzato dalla compagnia per effettuare il volo
- Distance: distanza percorsa in chilometri

Infine, l'ultimo dataset disponibile per la nostra analisi, è un file di tipo csv contenente i seguenti attributi:

- Country: nazione
- Country_Code: sigla della nazione
- City: città
- Airport_Name: nome dell'aeroporto
- Airport_Code: sigla dell'aeroporto

4.2. Correspondences Investigation

Durante la fase di *Correspondences Investigation*, si sono analizzati tutti gli attributi dei dataset a nostra disposizione per cercare di capire se fossero presenti delle corrispondenze.
Osservando il dataset delle rotte europee ottenuto con tecniche di web scraping, sono subito emerse le variabili *Departure* e *Arrivals* dal momento che contengono le sigle di aeroporti proprio come la variabile *Airport_ Code* del dataset in formato csv.
A questo punto dunque si è concluso che si possono unire queste due sorgenti di dati sulla base del codice univoco degli aeroporti. In seguito si è analizzato il dataset contenente le informazioni sulle compagnie aeree per trovare una corrispondenza con il dataset delle rotte. Come risultato è scaturito che entrambi contengono la variabile *Airline* che può essere utilizzata come corrispondenza per unire le due fonti di dati.

4.3. Data Preparation

Alla luce delle considerazione delle sezione precedente, si è svolta un'analisi sui dati contenuti negli schemi delle corrispondenze.
L'analisi non ha rilevato alcun problema sulla variabile *Airline*, ciò a significare dunque che c'è una corrispondenza perfetta tra il dataset delle rotte europee e il csv contente informazioni sugli aeroporti. Per quando concerne le sigle degli aeroporti, l'analisi è stata eseguita due volte, la prima riguarda la corrispondenza tra *Departure* (variabile del dataset delle rotte) e *Airport_ Code* (variabile del csv), mentre la seconda è stata eseguita tra la variabile *Arrivals* e *Airport_ Code*. Il risultato è che entrambe hanno prodotto valori mancanti, in totale 385, evidenziando così la presenza di alcuni mismatch o assenza di corrispondenza tra le due chiavi.

4.3.1 Valori mancanti

A seguito di alcune analisi, è emerso che il dataset di tipo csv non conteneva i dati di alcuni aeroporti europei, contenuti al contrario nel dataset delle rotte dal momento che essi sono stati operativi durante l'ultima settimana di Gennaio 2023.
Le informazioni degli aeroporti mancanti saranno elencate qui in seguito:

- ADA: Adana Airport, Turchia
- ADB: Izmir Adnam Menderes Airport, Turchia
- ALA: Almaty International Airport, Kazakhstan
- AYT: Antalya Airport, Turchia
- BJV: Milas-Bodrum Airport, Turchia
- DEB: Debrecen International Airport, Ungheria
- DLM: Dalaman Airport, Turchia
- ESB: Ankara Esenboğa Airport, Turchia
- IST: Istanbul Airport, Turchia
- PRN: Prishtina International Airport , Kosovo
- SAW: Istanbul Sabiha Gökçen International Airport, Turchia
- SKP: Skopje International Airport, Macedonia
- TZX: Trabzon Airport , Turchia
- BON: Bonaire International Airport, Netherlands
- BWF: Barrow/Walney Island Airport, United Kingdom
- CAT: Cascais Airport, Portugal
- CDT: Castellón–Costa Azahar Airport, Spain
- CFN: Donegal Airport, Ireland
- DEB: Debrecen International Airport, Hungary
- GNJ: Ganja International Airport, Azerbaijan

- HOQ: Hof–Plauen Airport, Germany
- IGS: Ingolstadt Manching Airport,Germany
- ILD: Lleida–Alguaire Airport, Spain
- LEU: Andorra–La Seu Urgell Airport, Andorra
- MLH: EuroAirport Basel Mulhouse Freiburg, Switzerland
- NAJ: Nakhchivan International Airport, Azerbaijan
- PRN: Prishtina International Airport, Kosovo
- QGY: Gyor Per Airport, Hungary
- QKT: Wevelgem Airport, Belgium
- SXM: Princess Juliana International Airport, Nederland
- XSW: Airbus Hamburg-Finkenwerder, Germany
- BXO: Airport Buochs, Switzerland

Il dataset contenente le informazioni sugli aeroporti, non contiene quelle relative alla Turchia, Kosovo, Macedonia e Andorra, mentre non contiene quelle relative ad alcuni aeroporti più piccoli e con meno traffico aereo. Per quanto riguarda l'aeroporto di Basilea *EuroAirport Basel Mulhouse Freiburg* la sigla contenuta nel file csv è BSL mentre nel dataset delle rotte è MLH. L'aereoporto di Ganja, Azerbaijan *Ganja International Airport*,la sigla contenuta nel csv è KVD mentre nel dataset delle rotte è GNJ.
Essendo nota la sigla degli aeroporti, sono state inserite manualmente le informazioni mancanti dal momento che sono state facilmente reperibili da diverse forti online, sono state poi aggiornate anche le sigle degli aeroporti uguali tra i dataset ma con sigle diverse.

4.4. Merge

Dopo le analisi è stato dunque eseguito un join di tipo Left sulle chiavi primarie menzionate nelle sezioni precedenti. L' integrazione del dataset delle rotte con il fle csv è avvenuta due volte:

- La prima volta si sono integrate le informazioni: *Country*, *Country_ Code*, *City* e *Airport_ Name* dell'aeroporto di **partenza** al dataset contenente le rotte
- La seconda volta si sono integrate le informazioni: *Country*, *Country_ Code*, *City* e *Airport_ Name* dell'aeroporto di **arrivo** al dataset contenente le rotte

L'integrazione con il dataset contenente le informazioni delle compagnie aeree europee (csv) è avvenuto una sola volta sullo schema *Airline*

Il dateset integrato cosi ottenuto ha dunque le seguenti colonne:

- Airline: compagnia aerea che ha effettuato il volo
- Airline_Code: sigla della compagnia aerea che ha effettuato il volo
- Departure: sigla dell'aeroporto di partenza della rotta
- Country_x: nazione di partenza della rotta
- Country_Code_x: sigla della nazione di partenza della rotta
- City_x: città di partenza della rotta
- Airport_Name_x: aeroporto di partenza della rotto
- Arrivals: sigla dell'aeroporto di arrivo della rotta
- Country_y: nazione di arrivo della rotta
- Country_Code_y: sigla della nazione di arrivo della rotta
- City_y: città di arrivo della rotta
- Airport_Name_y: nome dell'aeroporto di arrivo della rotta
- Aircraft: modello utilizzato nella rotta
- Distance: distanza percorsa dalla rotta

5. Database

In questa sezione del report verrà illustrata e motivata la scelta del nostro database, partendo con la definizione della tipologia utilizzata fino ad arrivare alla sua implementazione.
A seguire verranno eseguite alcune interrogazioni sul database per ottenere dei risultati di natura descrittiva, affiancati dall'esempio di alcune query utilizzare.

5.1. Schema Integration

Come già menzionato precedentemente, per lo svolgimento di questo progetto si è scelto di utilizzare un database di tipo NO-SQL data l'esigenza di andare oltre il modello relazione vista la nostra ridondanza tra dati. La tipologia di database NO-SQL scelta è quella di tipo Document Based dal momenti che consente di inserire i dati in uno schema ad albero. Infatti, i dati a nostra disposizione, vedono la migliore rappresentazione con questa tipologia di schema visto che sono presenti ridondanze tra attributi (*Country*, *Country_Code*, *City*, *Airport_Name*) riferite a due diverse variabili (*Departure*, *Arrivals*).
E' stato scelto dunque di utilizzare il database MongoDB a cause della sua capacità di memorizzare i dati in documenti flessibili JSON-like, il che significa che i campi possono variare da un documento all'altro ed è possibile modificare nel tempo la struttura dei dati.
I dati presenti nel nostro dataset integrato in formato csv sono stati inseriti nel database MongoDB attraverso l'utilizzo della libreria PyMongo di Python.
Il modello di dati scelto per l'implementazione, contiene due nidificazioni sulle variabili che contengono attributi ridondanti, così da poter facilmente maneggiare i dati nelle future interrogazioni del database senza rischio di ambiguità.
Esso ha la seguente forma:

```
{
  "_id": {
    "$oid": "63de9f256b9147b7fde5d869"
  },
  "Airline": "Aegean Airlines",
  "Airline_Code": "a3-aee",
  "Departure": {
    "Departure": "TIA",
    "Country_x": "Albania",
    "Country_Code_x": "AL",
    "City_x": "Tirana",
    "Airport_Name_x": "Tirana International Airport Nene Tereza"
  },
  "Arrivals": {
    "Arrivals": "ATH",
    "Country_y": "Greece",
    "Country_Code_y": "GR",
    "City_y": "Athens",
    "Airport_Name_y": "Athens International Airport"
  },
  "Aircraft": "32N",
  "Distance": "529"
}
```


Figura 5.1: Esempio di documento nel database MongoDB

5.2. Statistiche database

Attraverso la libreria PyMongo di Python sono state svolte alcune query sul nostro database di natura statistica.
Come primo passo si è svolta un'analisi sulle compagnie aeree per verificare quali fossero quelle che hanno eseguito più voli in Europa durante l'ultima settimana di Gennaio 2023.

Airline	Number of Fly
Ryanair	3190
Wizz Air	1145
easyJet	992
Vueling	428
Lufthansa	362
Eurowings	263
Jet2	236
British Airways	228
TUI Airways	214
Air France	187

Figura 5.2: Compagnie aeree con il maggior numero di rotte europee

Il risultato è che le compagnie aeree che hanno effettuato più voli sono quelle lowcost, tra cui al primo posto *Ryanair* con 3190 voli e a seguire *Wizz Air*, *EasyJet*, *Vueling*, *Lufthansa*, *Eurowings* e *Jet2*. Subito dopo di esse, le compagnie aeree con più voli sono quella di bandiera Britannica, TUI Airways (altra compagnia Britannica) e la compagnia di bandiera francese.
La query utilizzata per ottenere questo risultato è la seguente:

```
agg_result= db.airport.aggregate(
    [{
    "$group" :
        {"_id" : "$Airline",
         "total" : {"$sum" : 1}
         }}
    ])
```

Figura 5.3: Query per calcolare il numero di voli effettuati per ogni compagnia aerea

A seguito è stata svolta un'analisi sui chilometri totali percorsi dalle compagnie aeree con il risultato simile a quello precedente.

km	Airline
4493124	Ryanair
1715527	Wizz Air
1330670	easyJet
550612	Vueling
498151	Jet2
417474	TUI Airways
381068	Eurowings
379716	Lufthansa
245697	British Airways
220691	Transavia

Figura 5.4: Compagnie aeree con il maggior numero di km

Infatti le compagnie aeree che hanno percorso più chilometri sono *Ryanair* con 4.493.124 km nettamente superiore rispetto a tutte le altre, a seguire *Wizz Air*, *EasyJet* e *Vueling*
In generale, le compagnie aeree che hanno volato di più, sono quelle lowcost dal momento che consentono di viaggiare a prezzi molto ridotti e sono quindi accessibili a tutti. Entra tra le prime dieci però, la compagnia di bandiera Britannica che risulta aver totalizzato 245697 km nell'ultima settimana di Gennaio 2023, un risultato notevole anche se pur molto distanze da i chilometri percorsi dalle compagnie a basso costo menzionate precedentemente.

```
agg_result=db.airport.aggregate([{
  "$group": {
    "_id": "$Airline",
    "sum_distance": {
      "$sum":  {'$convert': { 'input': '$Distance', 'to': 'int' }}
    }
  }
},{
  "$project": {
    "Airline": "$_id",
    "sum_distance": 1,
    "_id": 0
  }
}])
```

Figura 5.5: Query per calcolare i chilometri percorsi da ogni compagnia aerea

Sorge spontaneo domandarsi però, se il risultato potrebbe variare drasticamente normalizzando i chilometri totali al numero di voli attuati da ogni compagnia aerea, ovvero di mettere in evidenza non solo la quantità di viaggi effettuati ma anche la loro lunghezza.
Infatti, svolgendo questa analisi, emerge che, le compagnie che hanno percorso più chilometri in rapporto al numero di viaggi, non sono quelle lowcost elencate precedentemente, ma bensì altre.
Esse verranno elencate qui in seguito.

Airline	Number of Fly	km_norm
Georgian Airways	6	2972.666667
Condor	72	2635.069444
Smartwings	5	2520.400000
Air Astana	2	2433.500000
SunExpress	53	2218.339623
Jet2	236	2110.809322
Icelandair	49	2074.000000
Azerbaijan Airlines	23	2060.826087
TUI Airways	214	1950.813084
Edelweiss Air	31	1902.258065

Figura 5.6: Compagnie aeree che hanno volato di più

Come si può vedere in Figura 5.6, la compagnia di bandiera Georgiana con soli sei voli ha percorso più chilometri di tutte le altre compagnie aeree in relazione al numero di voli effettuati. A seguire ci sono *Condor* compagnia aerea tedesca, *Smartwings* lowcost ceca e *Air Astana* compagnia di bandiera Kazaka.

Le analisi descrittive sono proseguite in seguito con un focus sugli aeroporti e sulle nazioni operative.

E' risultato che la nazione che ha percorso più volti è la Spagna con 1713 voli solo nell'ultima settimana di Gennaio 2023, a seguire Regno Unito e Italia

Country	Number of Fly
Spain	1713
United Kingdom	1426
Italy	1181
Germany	833
France	755
Poland	378
Portugal	374
Greece	307
Netherlands	294
Switzerland	266

Figura 5.7: Nazioni con più voli

```
agg_result= db.airport.aggregate(
    [{
    "$group" :
        {"_id" : "$Departure.Country_x",
         "total" : {"$sum" : 1}
         }}
    ])
```

Figura 5.8: Query per calcolare i numeri di voli percorsi da ogni nazione

Mantenendo il focus su queste tre nazioni, si è svolta un'analisi sulle mete più gettonate e le compagnie aeree più utilizzate per questi stati.

Spagna 570 voli percorsi su 1713 (33%) sono voli interni con partenza e arrivo da aeroporti spagnoli, 305 (18%) voli sono verso il Regno Unito, 163 (9.5%) verso la Germania, 132 (7.7%) verso l'Italia e 117 (6.8%) verso la Francia, l'elenco continua poi con altre minoranze.
Le compagnie aeree più utilizzate dagli spagnoli sono invece *Ryanair* con 568 (33%) voli percorsi, *Vueling* con 266 (15%) voli , *Iberia* 145 (8.5%) voli, *EasyJet* 122 (7.1%) voli e *Binter Canarias* con 73 voli (4,3%).

Regno Unito Per quanto riguarda il Regno Unito, solamente 206 voli su 1426 (14,4%) solo voli interni, mentre 306 voli sono diretti verso la Spagna (21.5%), 117 (8,2%) verso l'Italia, 108 (7.5%) verso la Francia e 88 (6,1%) verso la Polonia, l'elenco prosegue naturalmente con altre nazioni sempre più trascurabili.
Per quanto riguarda le compagnie aeree invece, la più utilizzata è *Ryanair* con 425 (39%) voli percorsi, a seguire *EasyJet* con 302 (21%) voli, *British Airways* con 125 (8.7%) voli, *Jet2* con 112 (7.8%) voli, *TUI Airways* con 100 (7%) voli, *Loganair* con 97 (6.8%) voli.

Italia Ben 346 voli su 1181 totali (31%) sono voli interni, ovvero con partenza e destinazione da aeroporti italiani, 132 (11%) voli sono diretti verso la Spagna, 117 (10%) verso il Regno Unito, 80 (6,7%) verso la Germania e cosa via con percentuali sempre minori verso altre nazioni europee.
Le compagnie aeree più utilizzate dagli italiani sono invece *Ryanair* con 599 voli percorsi su 1181 (50%), *Wizz Air* con 166 (14%) voli svolti, *EasyJet* con 102 voli (8,6%) e *ITA Airways* con 78 voli (6,6%)

Verranno ora mostrate le query utilizzate per ottenere questi risultati, come sempre è stata utilizzata la libreria Pymongo di Python. Per filtrare i risultati in base al nome della nazione (variabile *Country* è stato utilizzato l'operatore *Match* di MongoBD.

```
agg_result=db.airport.aggregate([{
  "$match": {
    "Departure.Country_x": "Italy"

  }
},{
  "$group": {
    "_id": "$Arrivals.Country_y",
      "count": {
      "$sum": 1
    }
  }
},{
  "$project": {
    "Arrivals": "$_id",
    "_id": 0,
      "count":1
  }
}])
```

Figura 5.9: Query per calcolare le destinazioni dei voli italiani

```
agg_result=db.airport.aggregate([{
  "$match": {
    "Departure.Country_x": "Italy"

  }
},{
  "$group": {
    "_id": "$Airline",
      "count": {
      "$sum": 1
    }
  }
},{
  "$project": {
    "Airline": "$_id",
    "_id": 0,
      "count":1
  }
}])
```

Figura 5.10: Query per calcolare le compagnie aeree utilizzate dagli italiani

Si può dedurre quindi che più del 30% dei voli effettuati dalla Spagna e dell'Italia sono voli interni dovuti magari al turismo nazionale, motivi di lavoro o ancora per studenti e lavoratori fuori sede che ritornano dalle loro famiglie.

Emerge inoltre che le nazioni estere preferite dagli spagnoli come meta di turismo o di lavoro sono Italia, Regno Unito e Germania, risultato molto simile per quanto riguarda gli italiani dal momento che le loro mete preferenze sono Spagna, Regno Unito e Germania. Inoltre spagnoli e italiani, hanno eseguito circa lo stesso numero di voli interni, superiore al 30%, ciò a significare un grande traffico aereo interno e un elevato numero di aeroporti nazionali a loro disposizione.

Risultati diversi si sono ottenuti per il Regno Unito, dal momento che le mete estere più gettonate sono Spagna, Italia, Francia e Polonia, con una percentuale di voli interni di solo 14.4%.

Per quanto concerne le compagnie aeree utilizzate, è risultato che la Spagna utilizza prevalentemente aerei di linea spagnola subito dopo *Rayanair*, risultato simile ottenuto anche dal Regno Unito dal momento che utilizza prevalentemente aerei di linea inglese. L'Italia invece, utilizza per lo più compagnie aeree lowcost come *Rayanair*, *Wizz Air* e *EasyJet* forse questo risultato è dovuto al fatto che non sono presenti compagnie lowcost italiane ma solo la compagnia di bandiera *ITA Airways* che rimane comunque la più scelta dopo le lowcost.

Le analisi descrittive sono proseguite con il calcolo degli aeroporti attivi durante l'ultima settimana di Gennaio 2023 per verificare se il primato andasse ancora una volta a Spagna, Regno Unito e Italia, così da motivare anche la grande quantità di voli da loro svolta.

```
agg_result=db.airport.aggregate([{
  "$group": {
    "_id": {
      "Airport_Code": "$Departure.Departure",
      "Country": "$Departure.Country_x"
    },
    "count": {
      "$sum": 1
    }
  }
},{
  "$project": {
    "Airport_Code": "$_id.Airport_Code",
    "Country": "$_id.Country",
    "count": 1,
    "_id": 0
  }
},{
  "$group": {
    "_id": "$Country",
    "count": {
      "$sum": 1
    }
  }
},{
  "$project": {
    "Country": "$_id",
    "count": 1,
    "_id": 0
  }
}])
```

Figura 5.11: Query per calcolare il numero di aeroporti attivi per ogni nazione

Il risultato è effettivamente quello atteso, dal momento che Regno Unito, Spagna e Italia sono tra le nazioni che hanno il maggior numero di aeroporti attivi, anche se il primato spetta alla Francia con ben 43 aeroporti. Per

	Airport	Country
0	43	France
1	42	United Kingdom
2	37	Greece
3	37	Spain
4	34	Italy
5	20	Germany
6	18	Sweden
7	17	Finland
8	13	Russian Federation
9	13	Poland

Figura 5.12: Nazioni con più aeroporti attivi

concludere, si è svolta un'analisi sugli aeroporti europei più attivi durante l'ultima settimana di Gennaio 2023, calcolando il numero di partenze per ognuno di essi.

Airport	Departure
MAD	199
AMS	198
BCN	191
DUB	172
VIE	165
MAN	163

Figura 5.13: Aeroporti più attivi

L'aeroporto con il più alto numero di partenze è quello di Madrid con ben 199 partenze, subito a seguire quello di Amsterdam con 198 partenze e quello di Barcellona con 191.

```
agg_result=db.airport.aggregate([
        {"$group" : {"_id":"$Departure.Departure", "count":{"$sum":1}}}
    ])
```

Figura 5.14: Query per calcolare il numero di voli da ogni aeroporto

6. Conclusioni e Sviluppi Futuri

Questo progetto è stato sviluppato eseguendo diverse operazioni di Data Management partendo dallo Web Scraping e dall Web API per l'estrazione dei dati fino ad arrivare all'implementazione e interrogazione di un database di tipo NO-SQL.
Questi passaggi ci hanno permesso di ottenere una grande base di dati completa e ben fornita di tutte le rotte aeree europee eseguite durante l'ultima settimana di Gennaio 2023.
Un possibile sviluppo futuro potrebbe essere sicuramente quello di integrare le informazioni nel nostro database con rotte aeree di altre settimane dell'anno, sarebbe ottimale infatti costruire un algoritmo in grado di eseguire lo scraping sui dati delle rotte quotidianamente.
Un altro sviluppo futuro consiste nell'ottenere dati di voli aerei internazionali e non solo europei così da avere una base di dati il più completa possibile.