Case Study: Airline on-time performance

Corso di Metodi Statistici per i Big Data - Prof. La Rocca Michele

Roberto Senatore

30 ottobre 2020

Introduzione

Questo breve report si riferisce all'analisi del dataset sulle performance degli aerei di linea negli Stati Uniti d'America, in particolare vuole mostrare, con diversi tipi di visualizzazione, le informazioni contenute nei dati, utilizzando strumenti per processare e computare problemi su una grande mole di dati, come Spark, e visualizzare i risultati con strumenti grafici nell'ambiente R.

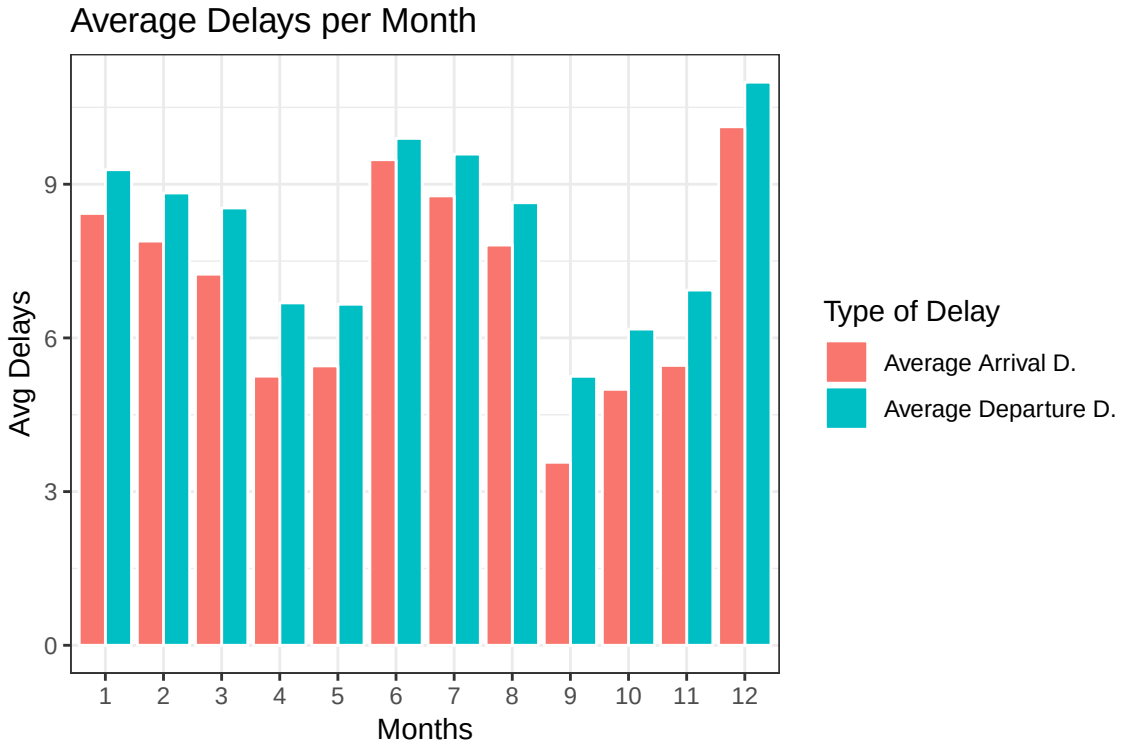
I dati

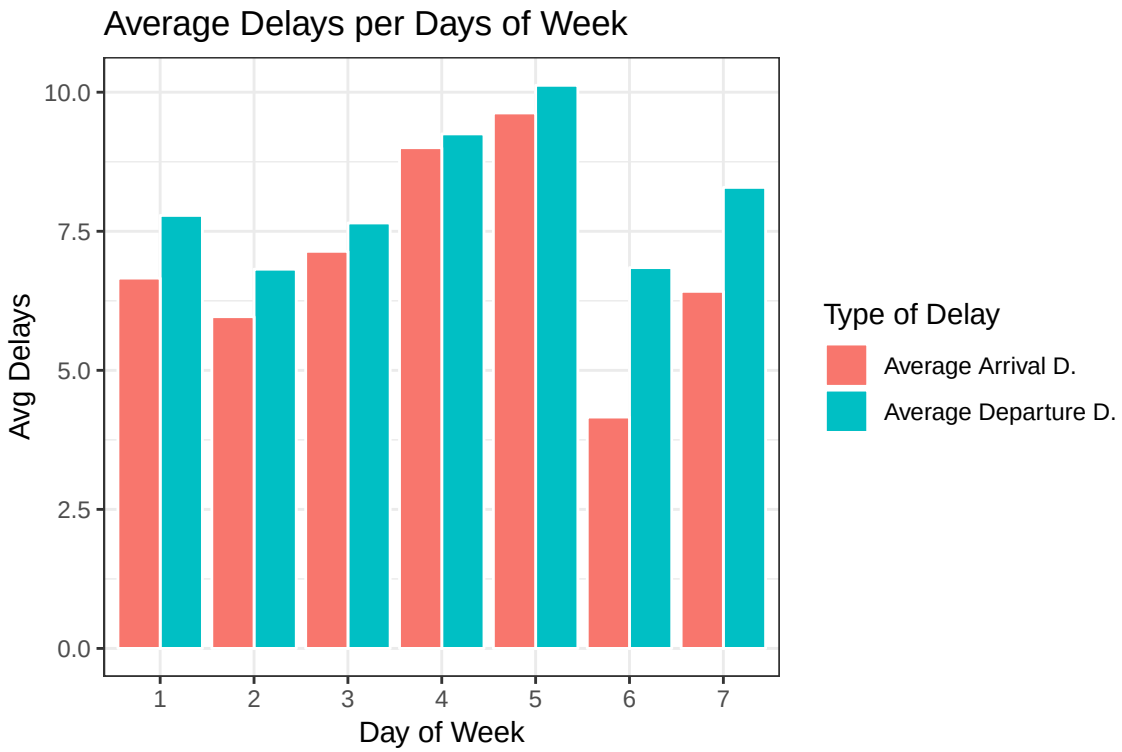
I dati sono stati scaricati dall'ASA e consistono sui voli di aerei di linea americani tra l'ottobre 1987 e Aprile 2008. Essi consistono di 120 milioni di osservazioni su 29 variabili, tuttavia non tutte sono state utilizzate.

Analisi proposte

Ritardi dei voli

La prima analisi si è rivolta sul problema più grande di quando si prende un aereo, cioè i ritardi. Si sono analizzati dapprima i migliori mesi dell'anno per volare per minimizzare i ritardi degli aerei e attraverso le variabili *ArrDel* e *DepDel* è stata calcolata la media dei ritardi di arrivo e di partenza degli aerei.

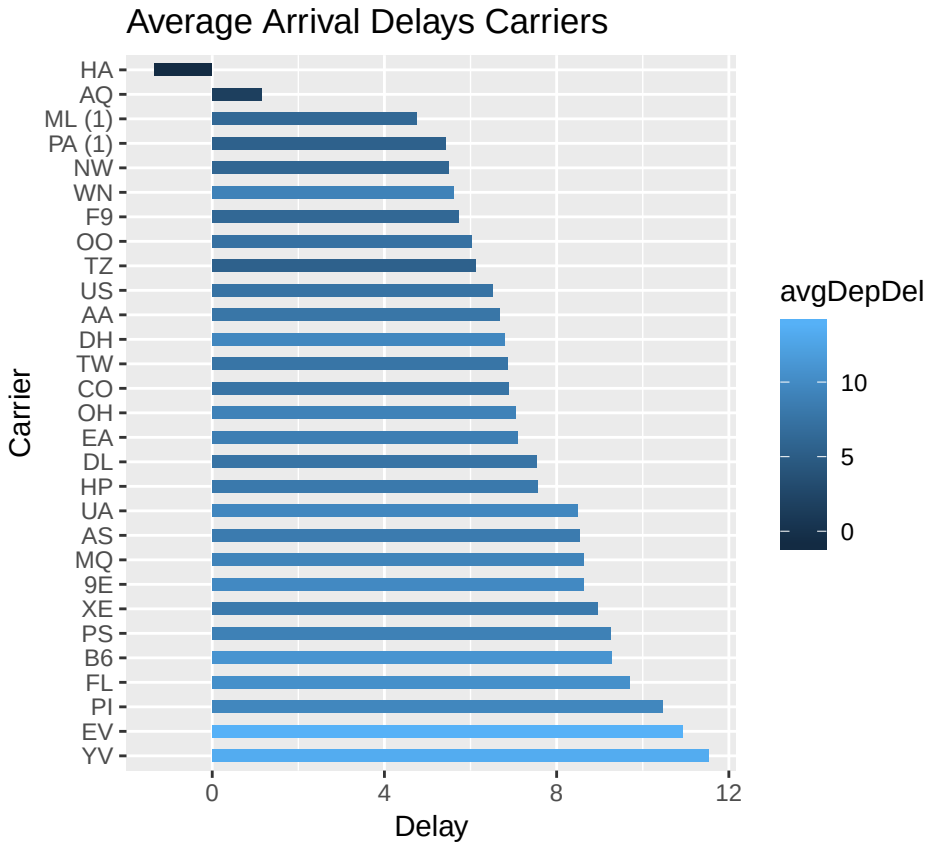


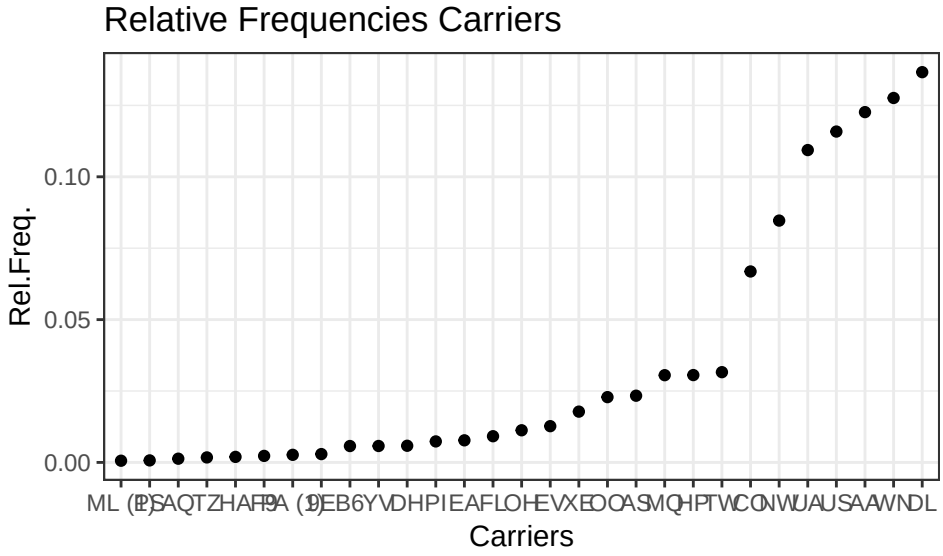


I grafici mostrano i ritardi medi di **arrivo** e di **partenza** degli aerei, le barre sono state riordinate in modo crescente. Il primo grafico mostra i mesi che minimizzano i ritardi e in cui è preferibile viaggiare, si può notare immediatamente come questi siano quelli di **Settembre-Ottobre** e quelli di **Aprile-Maggio** per entrambi i tipi di ritardi. Il secondo grafico, invece, mostra i migliori giorni per viaggiare che sono il **Venerdi** e il **Martedi**.

Compagnie aeree

La seconda analisi si è occupata di identificare le migliori e peggiori compagnie aeree in base ai ritardi, inoltre è stato utilizzato il dataset secondario *carriers* per associare i codici identificativi ai nomi esatti delle compagnie.





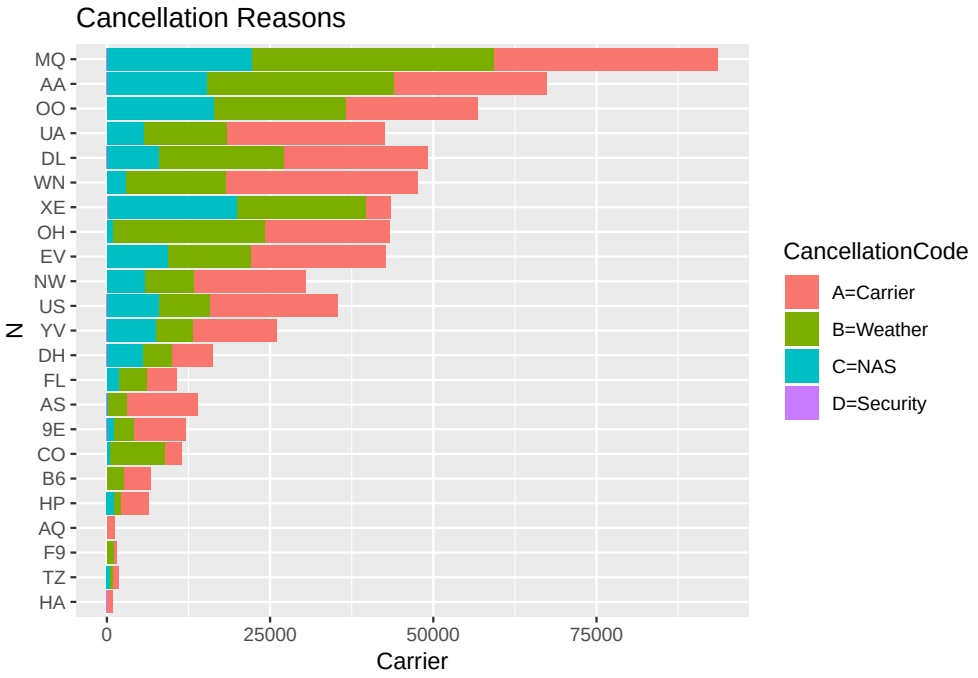
Il primo grafico rappresenta tutte le compagnie aeree ordinate dall'alto in basso rispetto il ritardo medio di arrivo, e in base al colore per il ritardo medio di partenza. La prima (HA) registra anticipi dei voli e la seconda (AQ) i tempi di ritardo più in basso in assoluto. Tuttavia, è stato registrato una grossa differenza in valore assoluto delle osservazioni dei voli, il secondo grafico rappresenta le frequenze relative delle osservazioni e l'evidente disomogeneità, ragion per cui per identificare le migliori e peggiori compagnie si è proceduto effettuando una media pesata dei ritardi medi rispetto il totale delle osservazioni. Tra le migliori ci sono: **Hawaiian Airlines**, **Aloha Airlines** e **Midway Airlines**. Tra le peggiori invece: **Delta Airlines**, **United**

Airlines e **Southwest Airlines**.

```
## # A tibble: 5 x 2
##   UniqueCarrier Description
##   <chr>         <fct>
## 1 HA            Hawaiian Airlines Inc.
## 2 AQ            Aloha Airlines Inc.
## 3 ML (1)        Midway Airlines Inc. (1)
## 4 PS            Pacific Southwest Airlines
## 5 TZ            ATA Airlines d/b/a ATA
```

```
## # A tibble: 5 x 2
##   UniqueCarrier Description
##   <chr>         <fct>
## 1 DL            Delta Air Lines Inc.
## 2 UA            United Air Lines Inc.
## 3 WN            Southwest Airlines Co.
## 4 AA            American Airlines Inc.
## 5 US            US Airways Inc. (Merged with America West 9/05. Reporting ~
```

Sono stati visualizzati anche i motivi di cancellazione dei voli, e quanti di questi siano imputabili direttamente alle compagnie (Cancellation Code: A):



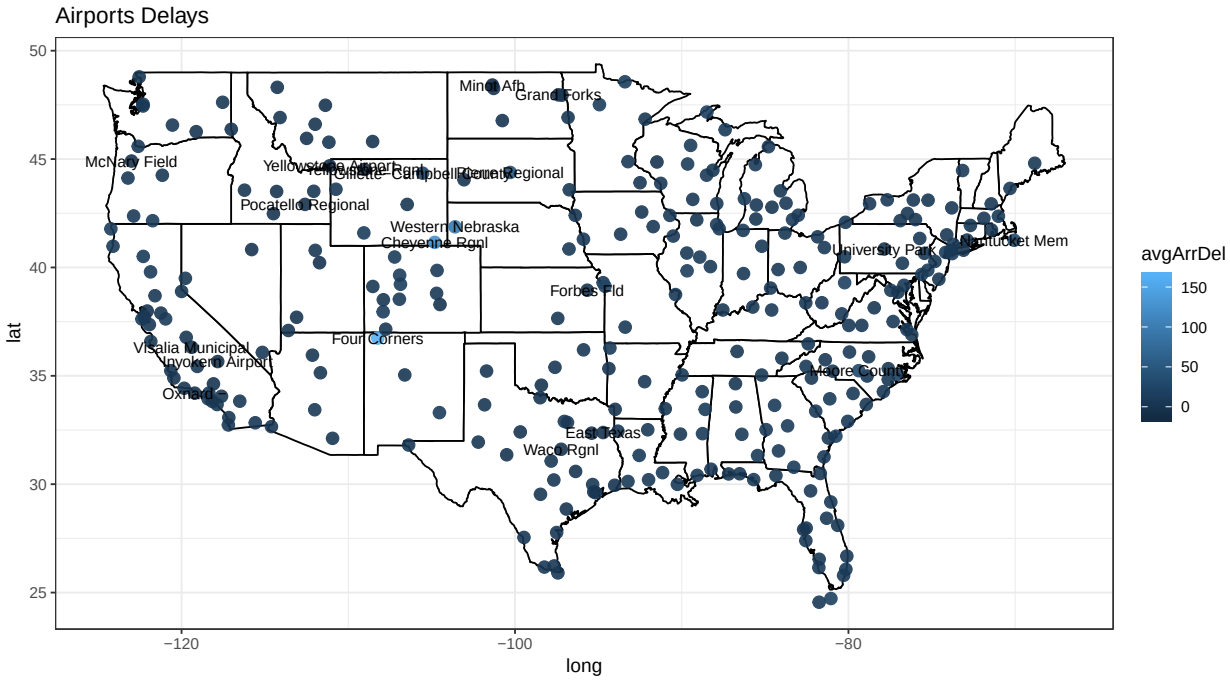
Aeroporti

Succesivamente, sono state analizzate anche le performance degli aeroporti in relazione ai ritardi dei voli in transito in essi. Anche qui, il dataset secondario *airports* è stato utilizzato per estrarre i nomi degli aeroporti. Anche qui sono state utilizzate delle medie pesate. Di seguito, l'elenco dei migliori e peggiori aeroporti e il grafico relativo sulla mappa degli Stati Uniti.

```
## # A tibble: 5 x 2
##   Origin name
##   <chr>  <chr>
## 1 MIB    Minot Afb
```

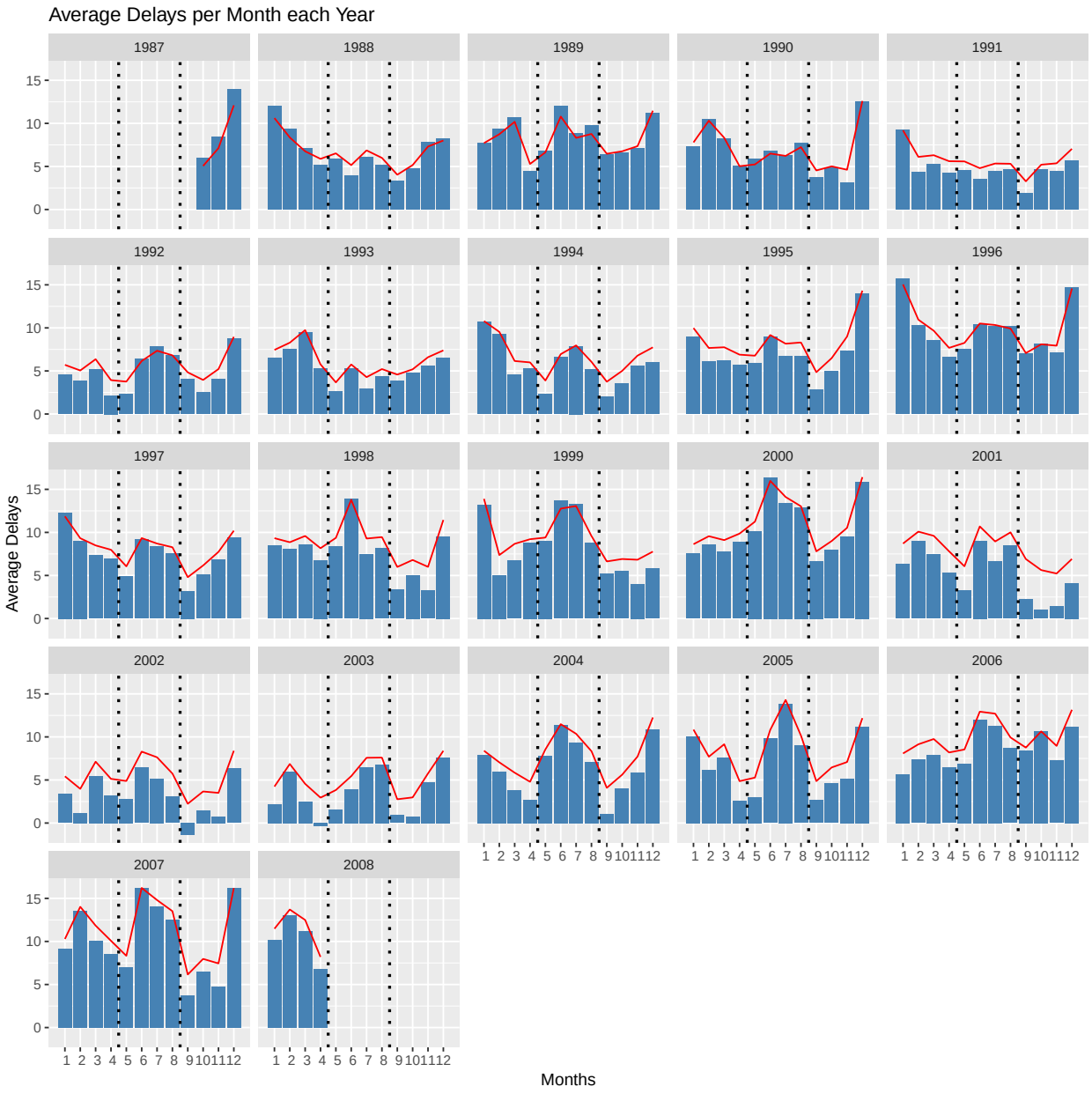
```
## 2 MKK    Molokai
## 3 LNY    Lanai
## 4 RDR    Grand Forks Afb
## 5 ROP    Rota International Airport
```

```
## # A tibble: 5 x 2
##   Origin name
##   <chr>  <chr>
## 1 FMN    Four Corners Rgnl
## 2 OGD    Ogden Hinckley Airport
## 3 CYS    Cheyenne Rgnl Jerry Olson Fld
## 4 BFF    Western Nebraska Regional Airport
## 5 PIR    Pierre Regional Airport
```



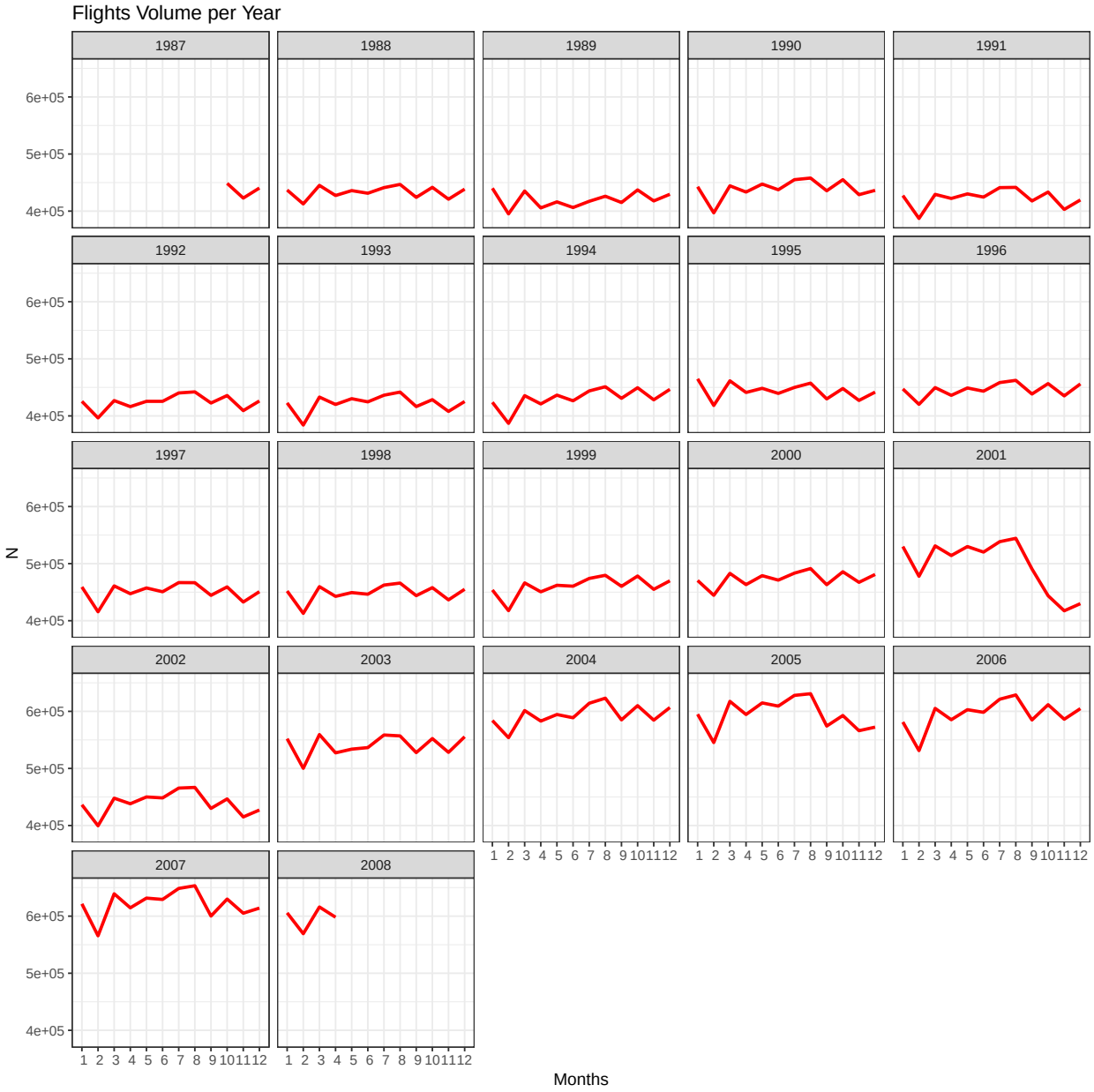
Fattori stagionali

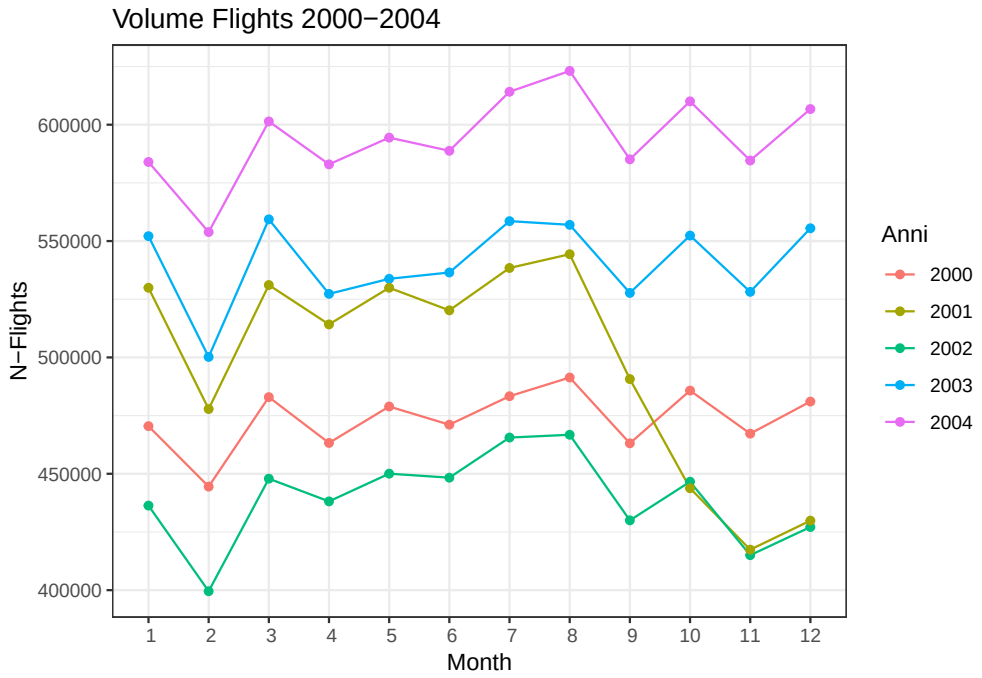
Nelle tantissime osservazioni si sono analizzati evenutali trend stagionali durante gli anni per quanto riguarda i ritardi dei voli. Sull'asse delle ascisse sono stati posti i mesi, divisi per quadrimestri, e sull'asse delle ordinate i ritardi di arrivo, con la linea rossa si evidenziano i ritardi delle partenze.

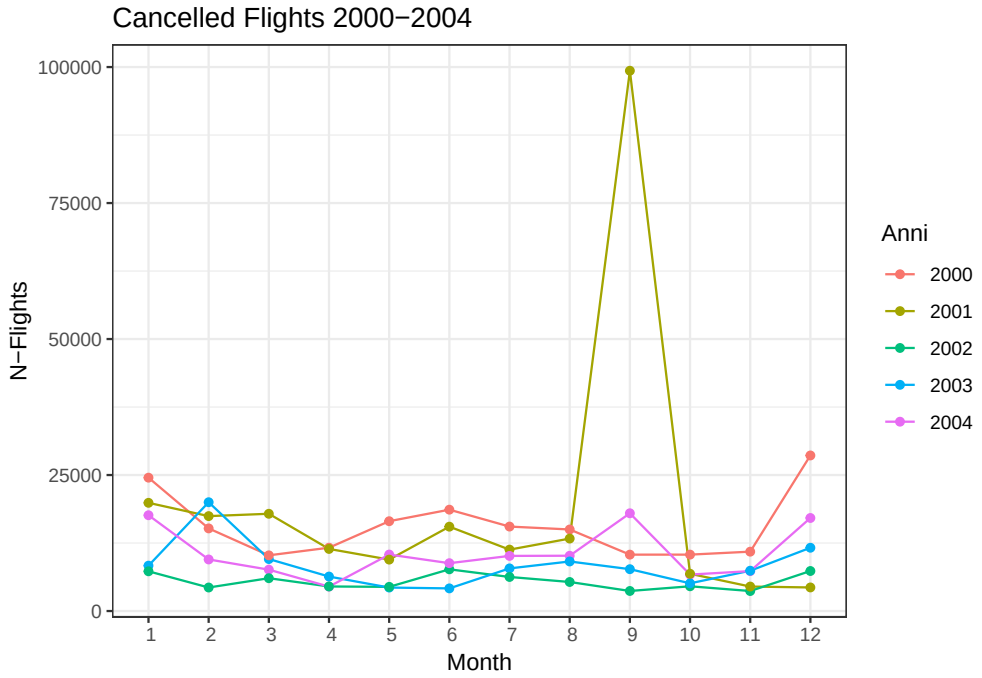


Volume dei voli

Infine, è stato analizzato come è cambiato il volume dei voli duranti gli anni, in cui si nota l'evidente crescita dagli anni 2000 in poi, e la sua relativa frenata con l'incidente del 9/11/2001, gli ultimi grafici analizzano più da vicino il cambiamento in quel periodo rispetto il volume e i voli cancellati.


Flights Volume per Year
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
N
6e+05
5e+05
4e+05
1 2 3 4 5 6 7 8 9 10 11 12
Months


Volume Flights 2000−2004
N−Flights
600000
550000
500000
450000
400000
1
2
3
4
5
6
7
8
9
10
11
12
Month
Anni
2000
2001
2002
2003
2004


Cancelled Flights 2000−2004
N−Flights
100000
75000
50000
25000
0
1
2
3
4
5
6
7
8
9
10
11
12
Month
Anni
2000
2001
2002
2003
2004