

{Progetto SQL}

Il rischio di viaggiare nel mondo per i turisti americani

Alessandro Smajlovic

I dati che analizzeremo per questo progetto
sono stati presi dal Dipartimento di Stato
degli U.S.A.

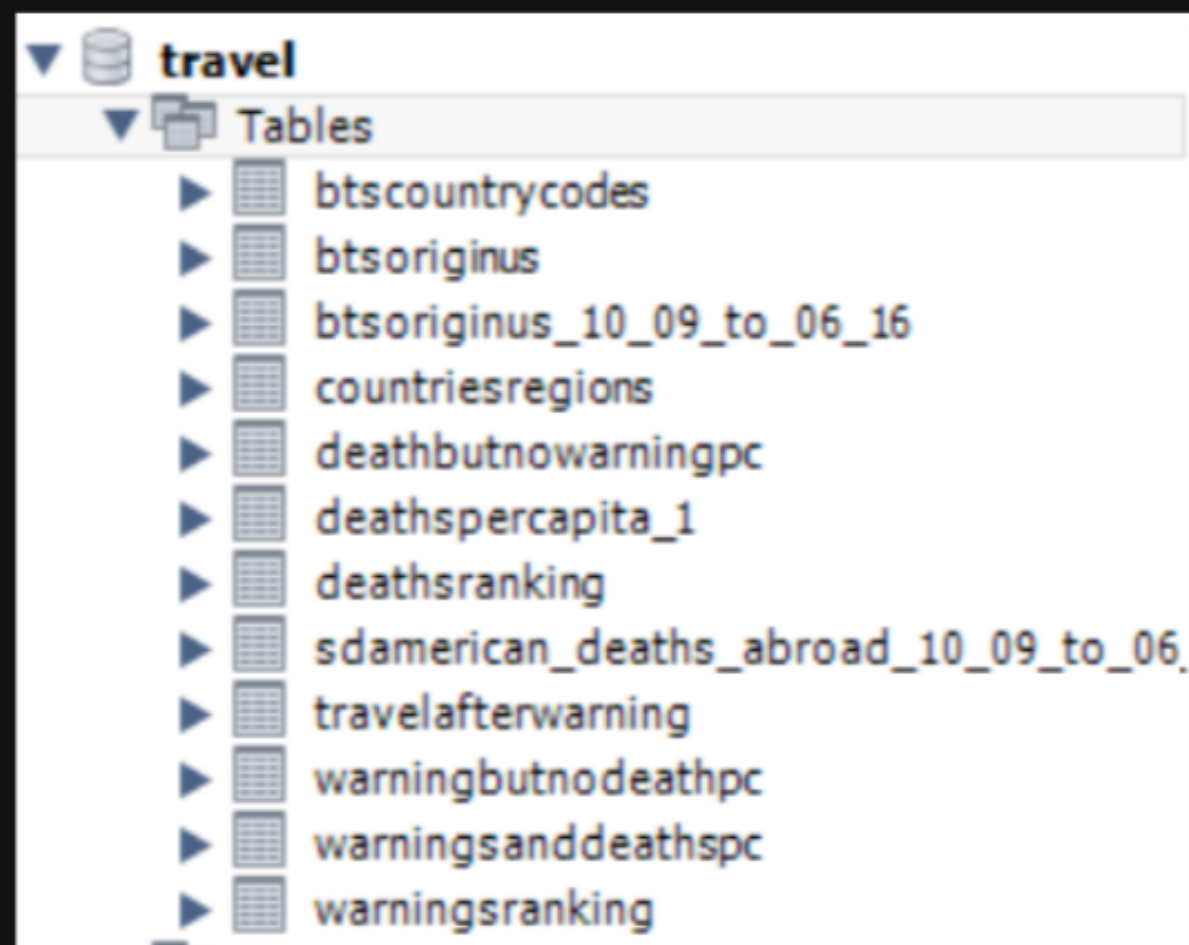


Progetto



L'obbiettivo è capire se esiste una relazione significativa tra il numero pro capite di decessi americani all'estero e il numero di segnalazioni che un Paese riceve. Messico, Mali e Israele sono stati indicati dalla maggior parte degli enti di segnalazione negli ultimi anni, ma è più probabile che dei turisti americani vengano uccisi in Thailandia, Pakistan e Filippine. Il compito era di cercare di fare chiarezza e capire davvero quali Stati hanno tassi di rischio più alti e se sono gli stessi indicati nel report. Infine si doveva comprendere quali sono quelli con tassi di rischio più bassi.



Il progetto prevede l'analisi su una serie di tabelle create dentro un database che ho nominato "travel" in MySQL, attraverso l'importazione dei file CSV(comma separated value) e messe in relazione tra loro grazie a una serie di joins, come avviene in un database relazionale.



In questo caso vado a vedere quali paesi hanno più vittime, prendendo tutte le colonne della tabella "travel.deathsraking" e vado a ordinarli per freq.

Quindi per prima cosa attraverso i select creiamo delle query che ci restituiranno come output una tabella

```
1 • SELECT *  
2 FROM travel.deathsraking  
3 ORDER BY freq DESC;
```


Result Grid   Filter Rows: <input type="text"/>			
	MyUnknownColumn	country	freq
▶	57	Mexico	598
	1	Afghanistan	84
	68	Philippines	74
	35	Haiti	65
	36	Honduras	46
	23	Dominican Republic	45
	43	Jamaica	39
	26	El Salvador	34
	21	Costa Rica	27
	33	Guatemala	26
	19	Colombia	25
	8	Belize	16
	24	Ecuador	12
	83	Thailand	11
	63	Nigeria	10
	4	Bahamas	9

Relazione significativa tra il numero pro capite di decessi americani all'estero e il numero di segnalazioni che un Paese riceve.

Ora per vedere se esiste una relazione significativa tra il numero pro capite di decessi e il numero di segnalazioni abbiamo usato una join. Prima di tutto siamo andati a rinominare le colonne, per poi utilizzare il left join, un operatore che ci restituisce tutti i record della tabella di sinistra (tabella 1) e i record corrispondenti della (tabella 2). Infine ho ordinato percap.

```
select distinct deathsperscapita_1.region as continente ,deathsperscapita_1.validctrys
as paese ,deathsperscapita_1.ntravelers as viaggi ,deathsperscapita_1.ndeaths as decessi,
warningsranking.SDwarnings_df_nWarnings as segnalazioni , percap FROM deathsperscapita_1
LEFT JOIN warningsranking
ON deathsperscapita_1.validctrys = warningsranking.SDwarnings_df_Country
ORDER BY percap DESC;
```

Qui riusciamo a vedere tramite questa tabella come ci sia una relazione tra loro, in quanto confrontando paesi come Pakistan e Messico, il primo ha un procapite di 3.54 su 25 segnalazioni, mentre il secondo ha un procapite di 0.84 su 28 segnalazioni.

Result Grid  Filter Rows: <input type="text"/> Export:  Wrap Cell Content: 						
	continente	paese	viaggi	decessi	segnalazioni	percap
▶	Asia	Pakistan	226200	8	25	3.54
	Asia	Thailand	343500	11	4	3.2
	Asia	Philippines	3240000	74	20	2.28
	Americas	Haiti	3316700	65	10	1.96
	Americas	Honduras	2766800	46	9	1.66
	Africa	Nigeria	780900	10	23	1.28
	Americas	Belize	1563400	16	0	1.02
	Americas	Guyana	722200	7	0	0.97
	Africa	Egypt	579800	5	11	0.86
	Americas	Mexico	71608500	598	28	0.84
	Americas	Guatemala	3799600	26	0	0.68
	Americas	El Salvador	5324400	34	9	0.64
	Europe	Greece	691800	3	0	0.43
	Asia	Jordan	775600	3	0	0.39
	Americas	Jamaica	10606600	39	0	0.37

Confronto segnalazioni e viaggi.

In seguito andiamo a capire come le segnalazioni possono influenzare i viaggi.

Per questo ho utilizzato il `select distinct` per rimuovere i risultati duplicati, in tale modo isoliamo i valori unici e creiamo un set di dati più "pulito" e coerente. Poi ho cambiato nome alle colonne che volevo riportare. Vado a unire le mie due tabelle d'interesse selezionando i dati con i valori comuni. Infine con `order by asc` sono andato a mettere i risultati per ordine ascendente.

```
select distinct warningsranking.sdwarnings_df_region as continente ,  
warningsranking.sdwarnings_df_country as paese ,  
warningsranking.sdwarnings_df_nwarnings as segnalazioni ,  
travelafterwarning.travelpctchange as viaggi  
from warningsranking  
inner join travelafterwarning  
on warningsranking.sdwarnings_df_country = travelafterwarning.travelcountries  
ORDER BY travelpctchange ASC;
```


Abbiamo ottenuto una tabella ordinata per viaggi e capiamo come il numero di segnalazioni non incida molto sui viaggi, ad esempio possiamo notare come il Messico nonostante il numero maggiore di segnalazioni ha un numero più alto di viaggi rispetto a un paese come l'Egitto che ha un risultato inferiore.

Result Grid					Filter Rows:	Export:	Wrap
	continente	paese	segnalazioni	viaggi			
►	Africa	Egypt	11	-34.0925385922439			
	Asia	Thailand	4	-14.9959205142651			
	Asia	Pakistan	25	-2.99854571315169			
	Asia	Philippines	20	-2.3403782211674			
	Americas	Venezuela	7	-1.40665509681445			
	Americas	Honduras	9	-1.17907014353109			
	Asia	Israel	25	-0.056097840906795			
	Americas	Mexico	28	0.515420158651302			
	Americas	Haiti	10	0.530475306340426			
	Americas	El Salvador	9	1.76424742785692			
	Europe	Russia	6	4.02583699248735			
	Americas	Colombia	18	5.96900134207724			
	Asia	Bahrain	3	6.55102395379668			
	Africa	Nigeria	23	10.9910652244429			
	Europe	Ukraine	15	11.3687615413101			
	Asia	Saudi Arabia	16	11.9810542016793			

Paesi più pericolosi per gli americani.

Messico, Mali e Israele sono stati indicati dalla maggior parte degli enti di segnalazione negli ultimi anni, ma è più probabile che dei turisti americani vengano uccisi in Thailandia, Pakistan e Filippine.

Attraverso questa query oltre inserire il left join, andremo ad analizzare e confrontare le nazioni specifiche di nostro interesse che vogliamo includere, per avere una visione semplificata del risultato.

```
• SELECT distinct deathspcapita_1.region AS continente,  
  deathspcapita_1.validctrys as paese,  
  deathspcapita_1.ntravelers as viaggiatori,  
  deathspcapita_1.ndeaths as decessi,  
  warningsranking.SDwarnings_df_nWarnings as segnalazioni, percap  
from deathspcapita_1  
left join warningsranking  
on deathspcapita_1.validctrys = warningsranking.SDwarnings_df_Country  
where SDwarnings_df_Country in ("Mexico", "Israel", "Mali", "Pakistan", "Philippines", "Thailand")  
order by SDwarnings_df_nWarnings DESC;
```

Il risultato ottenuto nella tabella è il seguente:
Pakistan, Filippine e Thailandia sono paesi più pericolosi rispetto a Messico e Israele.

Result Grid						
			Filter Rows:		Export:	Wrap Cell
	continente	paese	viaggiatori	decessi	segnalazioni	percap
▶	Americas	Mexico	71608500	598	28	0.84
	Asia	Israel	4770700	7	25	0.15
	Asia	Pakistan	226200	8	25	3.54
	Asia	Philippines	3240000	74	20	2.28
	Asia	Thailand	343500	11	4	3.2

Paesi più a rischio

Adesso analizzeremo i paesi più a rischio mostrando le morti in ogni paese, continente, numero di morti e viaggiatori, usando la clausola `left join` che ci restituisce sempre i records della tabella 1 (`deathsoercapita_1`) e i matching records della tabella 2 di destra (`warningsranking`). Infine ho filtrato con `Where` per poi ordinare per procapite.

```
• select distinct deathspercapita_1.region as continente,  
  deathspercapita_1.validctrys as paesi,  
  deathspercapita_1.ntravelers as viaggi,  
  deathspercapita_1.ndeaths as morti,  
  warningsranking.SDwarnings_df_nWarnings as segnalazioni,percap  
from deathspercapita_1  
left join warningsranking  
on deathspercapita_1.validctrys = warningsranking.SDwarnings_df_Country  
where percap >0 and ndeaths >1  
order by percap desc;
```

Il risultato come dimostra la tabella che nel continente Asiatico,specificando il paese del PAKISTAN nonostante il numero di viaggi inferiori rispetto ad altri paesi come Messico e Jamaica abbia una media procapite superiore

Result Grid						
Filter Rows:						
Export:						
Wrap Cell Content:						
	continente	paesi	viaggi	morti	segnalazioni	percap
▶	Asia	Pakistan	226200	8	25	3.54
	Asia	Thailand	343500	11	4	3.2
	Asia	Philippines	3240000	74	20	2.28
	Americas	Haiti	3316700	65	10	1.96
	Americas	Honduras	2766800	46	9	1.66
	Africa	Nigeria	780900	10	23	1.28
	Americas	Belize	1563400	16	0	1.02
	Americas	Guyana	722200	7	0	0.97
	Africa	Egypt	579800	5	11	0.86
	Americas	Mexico	71608500	598	28	0.84
	Americas	Guatemala	3799600	26	0	0.68
	Americas	El Salvador	5324400	34	9	0.64
	Europe	Greece	691800	3	0	0.43
	Asia	Jordan	775600	3	0	0.39
	Americas	Jamaica	10606600	39	0	0.37

Morti per terrorismo dal 10_09 al 06_16

Siamo andati a vedere la causa di morte per terrorismo nei confronti dei turisti americani, nell'arco del periodo menzionato.

Contando la causa di morte e rinominandola ho filtrato tramite where per poi usare l'operatore like che mi ha permesso di specificare un preciso carattere dalla colonna.

```
SELECT country, cause_of_death, COUNT(cause_of_death) as conteggio_causa_morte
FROM travel.sdamerican_deaths_abroad_10_09_to_06_16
where cause_of_death like 'Terrorist%'
GROUP BY country, cause_of_death
ORDER BY conteggio_causa_morte DESC;
```

Come si può notare il paese con più morti causa terrorismo risulta l'Afghanistan.

	country	cause_of_death	conteggio_causa_morte
▶	Afghanistan	Terrorist Action	73
	Jerusalem	Terrorist Action	5
	Belgium	Terrorist Action	4
	Somalia	Terrorist Action	4
	Albania	Terrorist Action	3
	Israel	Terrorist Action	3
	Iraq	Terrorist Action	2
	Jordan	Terrorist Action	2
	Syria	Terrorist Action	2
	Turkey	Terrorist Action	2
	Burkina Faso	Terrorist Action	1
	France	Terrorist Action	1
	Kenya	Terrorist Action	1
	Lebanon	Terrorist Action	1
	Mali	Terrorist Action	1
	Malta	Terrorist Action	1
	Uganda	Terrorist Action	1

Paesi con tassi di rischio più bassi.

Andiamo infine a vedere quali sono i paesi con tasso di rischio più bassi e quindi sicuri.

Per questo sono andato a unire due tabelle e rinominando le colonne che volevo riportare ho filtrato percap e ndeaths andando a usare il simbolo "<" per individuare tutti i paesi con un tasso minore se non nullo e ordinandolo per il numero di viaggi.

```
• select distinct deathspcapita_1.region as continente,  
  deathspcapita_1.validctrys as paesi,  
  deathspcapita_1.ntravelers as viaggi,  
  deathspcapita_1.ndeaths as morti,  
  warningsranking.SDwarnings_df_nWarnings as segnalazioni,percap  
from deathspcapita_1  
left join warningsranking  
on deathspcapita_1.validctrys = warningsranking.SDwarnings_df_Country  
where percap <1 and ndeaths <1  
order by viaggi desc;
```

Così siamo andati ad ottenere una lista dei paesi più sicuri dov'è possibile viaggiare senza problemi.

Result Grid						
		Filter Rows:	Export:		Wrap Cell Content:	
	continente	paesi	viaggi	morti	segnalazioni	percap
▶	Asia	South Korea	15234000	0	0	0
	NA	The Bahamas	8580100	0	0	0
	Asia	Hong Kong	8445400	0	0	0
	Asia	Taiwan	7026900	0	0	0
	Europe	Switzerland	6641100	0	0	0
	Americas	Argentina	4581100	0	0	0
	Americas	Aruba	4260900	0	0	0
	Oceania	New Zealand	2501400	0	0	0
	Europe	Iceland	2359000	0	0	0
	Europe	Denmark	2331800	0	0	0
	NA	Sint Maarten	2185100	0	0	0
	Americas	Bermuda	1978200	0	0	0
	Americas	Turks and C...	1949100	0	0	0
	Europe	Sweden	1651900	0	0	0

Result 2

Grazie per l'attenzione.

Rilascio di sotto il link con le query

https://docs.google.com/document/d/14ULXcqOcHFXCDeyQWjblbKsOw1caLXh3/edit?usp=drive_link&oid=111715975938215708651&rtpof=true&sd=true



start2impact
UNIVERSITY