Vademecum
Corso Power Query

Combinare tabelle separate in una unica

Miele Vangelisti

02/09/2024 – 09/09/2024

Nozioni di base

Uno dei classici problemi che si presenta a chi deve analizzare i dati è la combinazione di due tabelle di dati separate in una sola per alimentare una Tabella Pivot.

Mentre questo è facile per gli esperti di SQL, in Excel la via è quella di usare un CERCA.VERT() o una combinazione di INDEX(MATCH()), funzioni non particolarmente intuitive per leggere i dati da una tabella all'altra.

Power Query ha introdotto un altro metodo per combinare due tabelle insieme che non comportasse competenze di SQL, l'uso di delle complicate formule di Excel o la capacità di costruire una struttura di database relazionale.

Tipi di combinazioni di dati

Esistono diversi modi per unire i dati (come ben sa chi progetta database) purtroppo, ma questo fatto è stato in qualche modo nascosto agli utenti di Excel che hanno a disposizione poche formule come ad esempio la CERCA.VERT che compie solo la cosiddetta Left Outer Join.

In Power Query, tuttavia, ci sono ora molti tipi di combinazione (o Join) supportati attraverso la finestra di dialogo Merge di Query.

Questi tipi di Join possono permetterci di trovare non solo dati corrispondenti, ma anche dati non corrispondenti: qualcosa di molto importante per chiunque cerchi di far corrispondere o riconciliare dati presenti in più tabelle.

Tipi di combinazioni di dati

Consideriamo le due tabelle nell'immagine, i dati tra queste tabelle sono correlati, ma ci sono un paio di particolarità da sottolineare.
La tabella Piano dei Conti sulla destra fornisce un elenco distinto di tutti i conti solo quando si combinano le colonne (o campi) Conto e Reparto per generare l'identificatore univoco. Se guardi attentamente, vedrai che i valori nelle prime quattro righe della colonna Piano dei Conti si ripetono nelle quattro righe successive, quindi ci sono chiaramente dei duplicati. Allo stesso modo, le prime quattro righe della colonna Reparto contengono tutte il valore 150, e le quattro successive contengono il valore 250.
Se li combini insieme, tuttavia, c'è solo un'occorrenza per ciascuna (64010-150, 64020-150, 64010-250, ecc...). La **combinazione è denominata «chiave»**.
Lo stesso schema si ripete nella tabella Transazioni sulla sinistra.

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Tipi di combinazioni di dati

Con un'altra analisi delle tabelle si notano altre particolarità interessanti. Il Piano dei Conti è una tabella che contiene i conti su cui le transazioni POSSONO essere registrate. Quindi, se esiste un conto che non viene mai utilizzato, non è un problema.

Dall'altro lato, se una transazione viene registrata su una combinazione conto-reparto (la chiave condivisa) che non esiste, c'è un problema importante.

Questo problema non è limitato ai dati contabili. Esiste ogni volta che devi far corrispondere, confrontare o riconciliare due tabelle (Clienti vs Fatturato, Articoli vs Prezzi, ecc.).

Nelle pagine seguenti sono presentate le tecniche di combinazione delle due tabelle sulla base della chiave condivisa e le particolarità di ognuna.

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Left Outer Join

E’ il tipo di join predefinito e funziona restituendo tutte le righe della tabella di sinistra e qualsiasi record corrispondente dalla tabella di destra. Gli elementi nella tabella di destra senza una corrispondenza vengono ignorati.

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Conto	Reparto	Data	Importo	PianoDeiConti.Conto	PianoDeiConti.Reparto	PianoDeiConti.Nome
64010	150	15/12/2015 00:00	8975	64010	150	Mele
64020	150	15/12/2015 00:00	13708	64020	150	Pere
64030	150	15/12/2015 00:00	32555	64030	150	Arance
64010	250	15/12/2015 00:00	22752	64010	250	Pomodori
64030	250	15/12/2015 00:00	19733	64030	250	Pane
64040	250	15/12/2015 00:00	33438	64040	250	Pizza
64015	150	15/12/2015 00:00	34147			
64010	350	15/12/2015 00:00	45876			

Right Outer Join

Funziona restituendo tutte le righe della tabella di destra e qualsiasi record corrispondente dalla tabella di sinistra. Gli elementi nella tabella di sinistra senza una corrispondenza vengono ignorati.

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Conto	Reparto	Data	Importo	PianoDeiConti.Conto	PianoDeiConti.Reparto	PianoDeiConti.Nome
64010	150	15/12/2015 00:00	8975	64010	150	Mele
64020	150	15/12/2015 00:00	13708	64020	150	Pere
64030	150	15/12/2015 00:00	32555	64030	150	Arance
64010	250	15/12/2015 00:00	22752	64010	250	Pomodori
				64040	150	Susine
64030	250	15/12/2015 00:00	19733	64030	250	Pane
				64020	250	Insalata
64040	250	15/12/2015 00:00	33438	64040	250	Pizza

Full Outer Join

Funziona restituendo tutte le righe da entrambe le tabelle. In questo caso, nota che non solo hai i record che corrispondono tra le tabelle, ma hai anche tutti i risultati non corrispondenti che vengono esposti tramite il Left Outer Join, così come quelli esposti dal Right Outer Join

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Conto	Reparto	Data	Importo	PianoDeiConti.Conto	PianoDeiConti.Reparto	PianoDeiConti.Nome
64010	150	15/12/2015 00:00	8975	64010	150	Mele
64020	150	15/12/2015 00:00	13708	64020	150	Pere
64030	150	15/12/2015 00:00	32555	64030	150	Arance
64010	250	15/12/2015 00:00	22752	64010	250	Pomodori
				64040	150	Susine
64030	250	15/12/2015 00:00	19733	64030	250	Pane
				64020	250	Insalata
64040	250	15/12/2015 00:00	33438	64040	250	Pizza
64015	150	15/12/2015 00:00	34147			
64010	350	15/12/2015 00:00	45876			

Inner Join

L'Inner Join mostra solo le righe dove c'è una corrispondenza esatta tra le colonne selezionate in entrambe le tabelle. Tutte le righe che non hanno una corrispondenza vengono eliminate dal risultato.

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Conto	Reparto	Data	Importo	PianoDeiConti.Conto	PianoDeiConti.Reparto	PianoDeiConti.Nome
64010	150	15/12/2015 00:00	8975	64010	150	Mele
64020	150	15/12/2015 00:00	13708	64020	150	Pere
64030	150	15/12/2015 00:00	32555	64030	150	Arance
64010	250	15/12/2015 00:00	22752	64010	250	Pomodori
64030	250	15/12/2015 00:00	19733	64030	250	Pane
64040	250	15/12/2015 00:00	33438	64040	250	Pizza

Left Anti Join

Righe dalla tabella di sinistra che non hanno corrispondenza nella tabella di destra. Questo tipo di join è particolarmente utile quando si cercano eccezioni o discrepanze nei dati, permettendo di identificare rapidamente i record che esistono in una tabella ma non nell'altra.

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

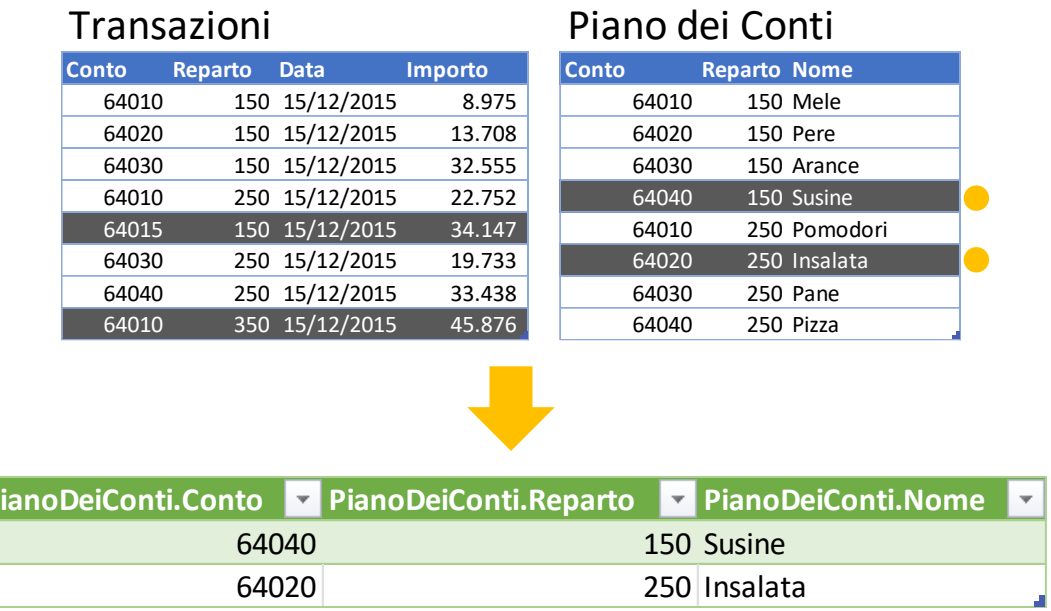
Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Conto	Reparto	Data	Importo
64015	150	15/12/2015 00:00	34147
64010	350	15/12/2015 00:00	45876

Right Anti Join

Righe della tabella di destra che non hanno corrispondenza nella tabella di sinistra. Tieni presente che ogni volta che crei un Right Anti Join, il risultato del join mostrerà una singola riga di valori nulli e una tabella annidata nell'ultima colonna. Se sei interessato solo agli elementi che non hanno corrispondenze, puoi fare clic con il tasto destro sulla colonna che contiene i risultati del merge e scegliere Rimuovi altre colonne prima di espanderla. È un modo efficace per trovare, ad esempio, conti che esistono nel piano dei conti ma che non sono stati utilizzati in alcuna transazione.



Full anti join

Tutte le righe che non hanno corrispondenze in entrambe le tabelle, è un tipo di join molto utile se stai cercando di identificare elementi che non corrispondono tra due tabelle. Questo non è un tipo di join predefinito disponibile attraverso l'interfaccia utente ma è facile creare la query: crea due query una Left Anti ed una Right Anti ed accodale.

Transazioni

Conto	Reparto	Data	Importo
64010	150	15/12/2015	8.975
64020	150	15/12/2015	13.708
64030	150	15/12/2015	32.555
64010	250	15/12/2015	22.752
64015	150	15/12/2015	34.147
64030	250	15/12/2015	19.733
64040	250	15/12/2015	33.438
64010	350	15/12/2015	45.876

Piano dei Conti

Conto	Reparto	Nome
64010	150	Mele
64020	150	Pere
64030	150	Arance
64040	150	Susine
64010	250	Pomodori
64020	250	Insalata
64030	250	Pane
64040	250	Pizza

Conto	Reparto	Data	Importo	PianoDeiConti.Conto	PianoDeiConti.Reparto	PianoDeiConti.Nome
64015	150	15/12/2015 00:00	34147			
64010	350	15/12/2015 00:00	45876			
				64040	150	Susine
				64020	250	Insalata

Prodotto Cartesiano (Cross Join)

Il **Prodotto Cartesiano** (noto anche come cross join o join molti a molti) consiste nel prendere i valori individuali da due tabelle e creare un insieme di coppie ordinate dove sono rappresentate tutte le possibili combinazioni.

La creazione di un Prodotto Cartesiano in Power Query può essere fatta andando su Aggiungi Colonna → Colonna Personalizzata dove nella formula, dopo l'uguale, va inserito il nome dell'altra tabella.

Da questo...

Month	Expenses	Budget
31/01/2024	Property Tax	5.450,00
29/02/2024	Insurance	1.253,00
31/03/2024	Telephony	1.640,00
30/04/2024		
31/05/2024		
30/06/2024		
31/07/2024		
31/08/2024		
30/09/2024		
31/10/2024		
30/11/2024		
31/12/2024		

A Questo...

Expenses	Budget	Month
Property Tax	5.450,00	31/01/2020
Property Tax	5.450,00	29/02/2020
Property Tax	5.450,00	31/03/2020
Property Tax	5.450,00	30/04/2020
Property Tax	5.450,00	31/05/2020
Property Tax	5.450,00	30/06/2020
Property Tax	5.450,00	31/07/2020
Property Tax	5.450,00	31/08/2020
Property Tax	5.450,00	30/09/2020
Property Tax	5.450,00	31/10/2020
Property Tax	5.450,00	30/11/2020
Property Tax	5.450,00	31/12/2020
Insurance	1.253,00	31/01/2020
Insurance	1.253,00	29/02/2020
Insurance	1.253,00	31/03/2020
Insurance	1.253,00	30/04/2020
Insurance	1.253,00	31/05/2020
Insurance	1.253,00	30/06/2020
Insurance	1.253,00	31/07/2020

...continua

Esercizio Prodotto Cartesiano

Trasformazioni eseguite nell'esercizio:

1. (ricavi) Rimuovi Colonne
2. (ricavi) Trasforma altre colonne tramite Unpivot
3. (ricavi) Rinomina colonne
4. (ricavi) Trasforma Tipo Dati in Numero Intero della colonna MeseRicavo
5. (dilazioni) Trasforma altre colonne tramite Unpivot
6. (dilazioni) Rinomina colonne
7. (ricavi) Aggiungi colonna personalizzata **(= dilazioni, effettua il Prodotto Cartesiano)**
8. (ricavi) Espandi colonna
9. (ricavi) Trasforma Tipo Dati delle colonne espanse
10. (ricavi) Aggiungi colonna personalizzata Incasso
11. (ricavi) Aggiungi colonna personalizzata MeseIncasso
12. (ricavi) Aggiungi colonna personalizzata AnnoBase
13. (ricavi) Aggiungi colonna condizionale Mese
14. (ricavi) Aggiungi colonna condizionale Anno
15. (ricavi) Aggiungi colonna personalizzata DataIncasso
16. (ricavi) Trasforma Tipo Dati in Data della colonna DataIncasso
17. (ricavi) Trasforma Data in ultimo giorno del mese
18. (ricavi) Rimuovi colonne non necessarie

Fuzzy matching

Ciascuno dei join illustrati finora in questo capitolo richiede che le chiavi condivise tra le due tabelle abbiano un'esatta corrispondenza.

Ma cosa succede quando si cerca di abbinare dati inseriti dall'uomo ai dati generati dal computer? Errori di ortografia, maiuscole/minuscole, abbreviazioni, punteggiatura e termini alternativi sono solo alcune delle cose che portano a incongruenze tra le tabelle da abbinare. Poiché i join predefiniti di Power Query uniscono solo i dati che corrispondono esattamente, questo può influire significativamente sulla capacità di confrontare due elenchi.

Le strategie sono due: 1) si torna indietro e si pulisce manualmente una delle due tabelle; 2) si **utilizza il metodo Fuzzy Matching per gestire l'ambiguità**.

Power Query sfrutta l'*algoritmo di similarità di Jaccard* per misurare la somiglianza tra coppie di elementi, e contrassegna tutto ciò che ottiene un punteggio di similarità dell'80% o superiore come una corrispondenza. In questo caso, l'algoritmo valuta sia "Laptops" oppure "laptop" come sufficientemente simili a "Laptop", nonostante uno abbia un carattere in più e l'altro usi un carattere minuscolo invece che maiuscolo. Parimenti anche caratteri trasposti (friend vs freind) e punteggiatura leggera (mrs vs mrs.) saranno abbinati laddove un join standard non lo farebbe.