Procedura esercizi W4D4 - M2 Excel

Esercitazione Finale

Ci tengo ad ammettere che molte applicazioni che andrò ad elencare sono soprattutto frutto di una condivisione e ragionamenti di gruppo, in questi 3 giorni mi sono davvero divertito a lavorare su questi esercizi, applicandomi prima di tutto da solo e inevitabilmente anche con molti dei miei colleghi di corso (principalmente tramite Discord), i quali hanno contribuito con spunti davvero interessanti. Credo di aver consolidato determinati argomenti più in questi tre giorni grazie al ragionamento di gruppo, alla condivisione di nozioni con i miei colleghi che in tutto il mese appena trascorso. Ho apprezzato molto la loro partecipazione e quindi li ringrazio! Spero di aver contribuito anche io nel mio piccolo.

Per l’esercizio 1) ho iniziato con un’analisi visiva del dataset, rendendomi conto dei molteplici dati “sporchi” presenti. Ho preferito provvedere subito alla pulizia del dataset utilizzando strumenti di Power Query e di Excel, così da ottenere nella tabella dinamica del foglio “RICERCA” informazioni coerenti. Ammetto di non essere riuscito, almeno in un primo momento, a correggere e pulire tutti i dati servendomi esclusivamente di Power Query.

* Ad esempio, ho sostituito i NONvalori delle tantissime celle vuote presenti nel dataset, con il testo “sconosciuto” per Indirizzo e con il testo “N.D.” utilizzato per le celle vuote di “indirizzo di posta elettronica” e “indirizzo internet”
* Nella colonna “Denominazione”, alcuni nomi di strutture avevano degli spazi davanti al nome, ho provveduto alla pulizia con una formula =annulla.spazi (avrei potuto fare la stessa cosa da Power Query con lo strumento “formato” taglia e pulisci); oltre alla cancellazione di alcuni doppioni di denominazioni strutture (utilizzando sia lo strumento “rimuovi duplicati” su excel o lo strumento rimuovi righe duplicate in Power Query). Mi sono reso comunque conto (attraverso un’analisi visiva e all’applicazione di una formula =CONTA.SE inserita nella colonna B, poi nascosta) che lo strumento “rimuovi duplicati” non aveva riconosciuto due righe con valori completamente uguali se non per la differenza del dominio della posta elettronica (.it e .com); in questo caso ho optato per una rimozione manuale delle righe doppie, solo dopo aver spostato nella cella l’indirizzo internet differente, così da avere per quella specifica struttura, due differenti indirizzi di posta elettronica. TOTALE RIMOZIONI FINALI 7; (P.S. potrei aver ragionato male in questi due casi, dal momento che forse le due strutture erano effettivamente due distinte e non gemelle alle altre due presenti nel dataset ma ho comunque valutato che fosse meglio rimuoverle, mantenendo cmq salva l’informazione differente (ovvero l'indirizzo di posta elettronica, accorpandolo all'altro già presente con il demilitatore /)..
* Per quanto riguarda la colonna “Indirizzo”, mi sono accorto (filtrando ad esempio per 1 “spazio” 1) che moltissimi civici davano un problema di ripetizione; pertanto, ho risolto a modo mio, creando in Power Query una colonna aggiuntiva con lo strumento “dividi colonna” – per delimitatore – spazio – da destra, una volta caricato in Excel ho provveduto a formattare la colonna indirizzi, riempiendo per colore le celle che non contenevano testo (ovvero tutte le celle alle quali avevo rimosso il civico), così facendo ho continuato a filtrare per valore contenente 1 (e così via fino al numero 9), in questo modo ho controllato circa 600 righe (errate) su circa 6.000 righe dell’intero dataset; Avrei potuto fare la stessa operazione rimanendo su Power Query, selezionando la colonna indirizzo e utilizzando lo strumento “formato” – taglia e pulisci ma ho notato che comunque molti errori persistevano (e quindi avrei comunque dovuto agire manualmente); quindi ho optato per una procedura di pulizia più macchinosa ma visivamente più comoda per me (ovvero quella sopra descritta, utilizzando Excel); P.S. qui ho un dubbio, ho provato a sistemare la cosa rimanendo su Power Query per mezzo dello strumento “colonna esempi”, l’intenzione era quella di digitare l’indirizzo corretto per una serie di campi in cui il civico era doppio (quindi digitandolo senza ripetere il civico doppio, sperando capisse la regola), il fatto è che per la colonna esempi, Power Query mi consentiva di digitare solo su in un range di 100 campi/righe, così facendo però (essendo gli errori sparsi su oltre 6.000 righe) i campi da me digitati non erano sufficienti affinché la colonna esempi capisse le modifiche da apportare (ho provato anche a cambiare l’ordinamento da crescente a decrescente), nella speranza che in quel range di 100 righe ci fossero abbastanza civici doppi da poter ridigitare affinché la correzione funzionasse ma i campi con i doppioni erano sempre inferiori ad es. a 5.. e quindi le modifiche avvenivano solo parzialmente se non in modo addirittura errato. Quindi, la mia domanda è, avevo modo di aumentare il range di caratteri da 100 a 1.000 ad esempio, così da poter digitare abbastanza esempi per fargli capire dove dover intervenire per rimuovere il civico duplicato?
* In merito alla colonna “Località”, ho notato che molte celle erano vuote, quindi anziché sostituire i valori delle celle “vuote” con il testo “sconosciuto” utilizzato per Indirizzo o con il termine “N.D.” utilizzato per le celle vuote di “indirizzo di posta elettronica” e “indirizzo internet” (come sopra descritto), ho deciso di inserire la “Città” in tutte quelle celle dove la località non era presente. P.S. ho ragionato così perché nel dataset grezzo (l'originale), nella maggioranza delle celle contenenti la località, in realtà il valore era proprio il testo della città di appartenenza; quindi, per dare una forma più ordinata possibile al dataset e mantenendo comunque la congruenza delle celle della colonna “Località” già compilate, ho agito così.
* Per quanto riguarda le colonne “indirizzo di posta elettronica” e di “indirizzo internet”, impostando un filtro e ricercando rispettivamente www e @, mi sono reso conto che alcuni valori erano invertiti, ovvero: alcune mail si trovavano nella colonna “indirizzo internet” e viceversa. Inizialmente, aiutandomi con laformattazione delle celle per colore, ho semplicemente trascinato la selezione (in blocco) nella colonna corretta per entrambi i casi. Ho potuto agire così in maniera sempliciotta, visto che i record sbaglliati erano circa un centinaio per la colonna "indirizzo di posta elettronica" e un centinaio per la colonna "indirizzo internet" e non solo erano per la maggior parte in righe sequenziali. In un secondo momento e grazie agli spunti forniti dai miei colleghi, ho voluto mettere in pratica altre possibili applicazioni, ad esempio, creando una colonna condizionale dove: per la colonna “indirizzo internet” se conteneva il valore della @ allora il valore andava spostato nella colonna “indirizzo di posta elettronica”; ho effettuato lo stesso procedimento per la colonna di posta. Alcune celle continuavano a mantenere errori, non filtrabili semplicemente per “www” e per “@”, quindi ho concluso la pulizia manualmente da Excel, applicando altri filtri e andando per tentativi. Avrei anche potuto suddividere in più fogli excel contenti i vari filtri applicati (per www o per @), ad esempio creando un foglio con tutte righe e colonne contenti le mail sbagliate (e il resto delle righe cancellarle), fare la stessa cosa su un altro foglio ma questa volta mantenendo le righe e colonne corrette, caricare il tutto in Power Query e una volta rinominate le colonne (ad esempio la tabella sbagliate in “corrette” e la tabella corrette in “sbagliate” e dopo aver invertito la loro posizione avrei potuto utilizzare lo strumento “accoda query come nuova”
* Nota: così facendo, ho potuto fornire alla tabella dinamica del foglio “RICERCA” delle informazioni utili e coerenti da restituire prima di arrivare a dover svolgere l'esercizio 3 per mettere in pratica la pulizia del dataset.

Foglio “RICERCA”:

* Per il menù a tendina color arancione ho usato lo strumento "convalida dati" – per elenco – dalla colonna di origine (precedentemente creata e poi nascosta), ovvero la colonna che concatena i dati di “Denominazione” e di “Città”. Accertandomi, per la nuova colonna concatenata, del fatto che non ci fossero stesse denominazioni di struttura per stessa città (attraverso lo strumento rimuovi duplicati per la colonna in questione); così facendo ho deciso di non utilizzare un codice identificativo per tutte quelle denominazioni che si ripetono seppur per città diverse, ad Es. “AREA DI SOSTA CAMPER”). Queste ripetizioni compromettevano l’utilizzo della tabella di Ricerca, in quanto, digitando “AREA DI SOSTA CAMPER” il menù a tendina mi restituiva un solo record per la prima città in ordine alfabetico. Quindi, credo di aver risolto accorpando la denominazione con la città. P.S. effettivamente così facendo sono andato ad inficiare sulla funzionalità della tabella. La cella B4 del foglio di “RICERCA”, mi restituisce una ridondanza di informazioni, restituendomi il nome della città (la quale mi viene già comunicata con l’inserimento della struttura nella cella B3)… ho provato ad ovviare a questa cosa, cercando di creare una seconda tendina (sempre nella tabella dinamica “RICERCA” anche per la cella B4 “CITTA” subordinata alla prima tendina, la quale si sarebbe dovuta aprire solo nel caso in cui avessi digitato una denominazione di struttura che si ripeteva, appunto “AREA DI SOSTA CAMPER” ad esempio, ma anche cercando sul web questa funzionalità, non sono riuscito ad applicarla..
* Per le celle azzurre (B4, B5, B6, B7) ho utilizzato una funzione =CERCA.VERT;
* Per le celle grige ho utilizzato un =CONTA.VALORI per la cella B8; un =CONTA.SE per la cella B9. P.S. l’esercizio chiedeva il numero di strutture per città ma in tabella era riportato per località, ho preferito modificare inserendo “per CITTA’”;

Per l’esercizio 2) ho inserito la tabella Pivot, mostrando il numero di strutture filtrabile per città come richiesto ma ho utilizzato un filtro dinamico anziché inserire il campo città in “filtri”;

Per l’esercizio 3) non ho dovuto apportare eccessive modifiche (se non per la tabella prezzi), in quanto avevo già pulito precedentemente il dataset per una questione di comodità; pertanto, ho creato un nuovo file chiamato MODELLO DATI.XLSX, caricando al suo interno (e trasformando con Power Query) le tabelle: “Strutture ricettive” dal file “elencostrutture” e “Prezzi” del file “prezzimedi”. Una volta apportate piccole correzioni conclusive, ho caricato entrambe le tabelle nel modello dati così da poterle relazionare in Power Pivot per mezzo della chiave univoca “Città” della tabella “Prezzi” verso il campo “Città” della tabella “Strutture ricettive”.

Anche se non richiesto dall’esercizio ho comunque generato una tabella Pivot conclusiva per avere un’applicazione anche visiva della correlazione tra le tabelle per mezzo della colonna Città.

Davide Buccino

20/01/2025