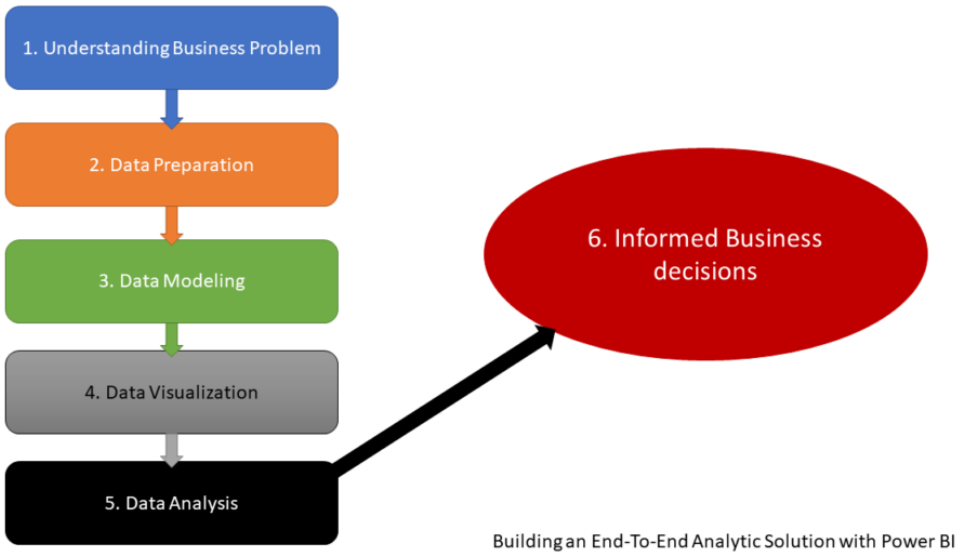
5A- SOLUZIONE COMPLETA - Prima Parte

Power BI: Parte 1 – Comprensione del problema aziendale!

In questo esercizio utilizzerò un set di dati aperto che contiene dati sugli incidenti automobilistici a New York e può essere trovato qui . Questo set di dati contiene circa 1,8 milioni di righe. Ogni riga rappresenta un incidente avvenuto a New York, dove almeno una persona è rimasta ferita/uccisa, o il danno complessivo è stato di almeno 1000$. I dati arrivano nel file CSV, contenente 29 colonne.

Ora, prima di iniziare a costruire la nostra soluzione, dobbiamo definire il flusso di lavoro e identificare le fasi specifiche del processo. Quindi, il primo e più importante compito è impostare i passaggi necessari per creare un risultato finale. Ecco la mia lista:

1. ***Comprendere il problema aziendale*** : questo è il punto di partenza, poiché senza comprendere il problema aziendale, la nostra soluzione non sarà in grado di soddisfare le esigenze aziendali. Voglio aumentare le vendite? La fidelizzazione dei clienti è il mio obiettivo principale? Cosa succede se scarto alcuni servizi nel prossimo trimestre? Questi sono alcuni esempi tipici delle domande aziendali a cui è necessario rispondere utilizzando le informazioni sui dati. In questo esempio, il nostro problema "aziendale" è identificare le posizioni critiche per le collisioni e cercare di prevenire incidenti in futuro
2. ***Preparazione dei dati*** : in questa fase, dobbiamo eseguire alcuni passaggi per preparare i nostri dati per un'ulteriore ingestione. A partire dalla profilazione dei dati , in modo da poter identificare possibili outlier e anomalie, quindi applicare variet ecniche di modellazione dei dati per preparare i dati PRIMA che diventino parte del nostro modello di dati
3. ***Modellazione dei dati*** – Poiché stiamo costruendo una soluzione analitica, il modello dei dati deve soddisfare (o almeno DOVREBBE soddisfare) alcuni postulati generali relativi alla modellazione dei dati. Per la maggior parte dei sistemi analitici, incluso Power BI, la modellazione dimensionale è la strada da percorrere, quindi dobbiamo scomporre la nostra tabella dei fatti originale e sfruttare il concetto di schema Star per stabilire il modello di dati corretto
4. ***Visualizzazione dei dati*** : questa è la fase che piacerà di più alle persone dall'inizio. È ora di soddisfare i nostri occhi con i numeri e di visualizzarli utilizzando i pratici elementi visivi di Power BI
5. ***Analisi dei dati*** : avere una bella visuale va bene, ma deve fornire alcune informazioni a una persona che la guarda. Pertanto, lo scopo principale di questa fase è fornire informazioni: ad esempio, quali sono le ore di punta per gli incidenti stradali a New York? Quali sono i luoghi più rischiosi? Quanti pedoni sono rimasti feriti nel Queens? E così via...
6. ***Decisioni aziendali informate*** : questa è una fase facoltativa e avrebbe potuto essere esclusa da questa soluzione e lasciata completamente agli stakeholder aziendali. Ma, ehi, svolgiamo il nostro ruolo di Data Analyst fino alla fine e diamo alcuni consigli basati sulle intuizioni che abbiamo ottenuto nella fase precedente!



Building an End-To-End Analytic Solution with Power BI

1. Comprensione del problema aziendale

Il primo e più importante passo per costruire la tua soluzione analitica (di successo), al fine di servire il suo scopo ed ***essere adottata dagli utenti*** , è dare risposte a domande aziendali chiave. Nessuno ha bisogno di dashboard carini e immagini interessanti se non forniscono informazioni e aiutano i decisori a capire cosa sta succedendo e perché.

Come posso aumentare le mie vendite? Perché così tanti clienti ci hanno lasciato nel trimestre precedente? Cosa posso fare per migliorare il processo di consegna? Qual è il periodo migliore per indirizzare il mercato con le promozioni?

Queste sono solo alcune delle domande più frequenti poste dagli stakeholder aziendali. Non solo: forse un'analisi dei dati sottostanti può aiutare gli utenti a identificare modelli completamente nuovi e porre una domanda: ***stiamo risolvendo il problema giusto?***

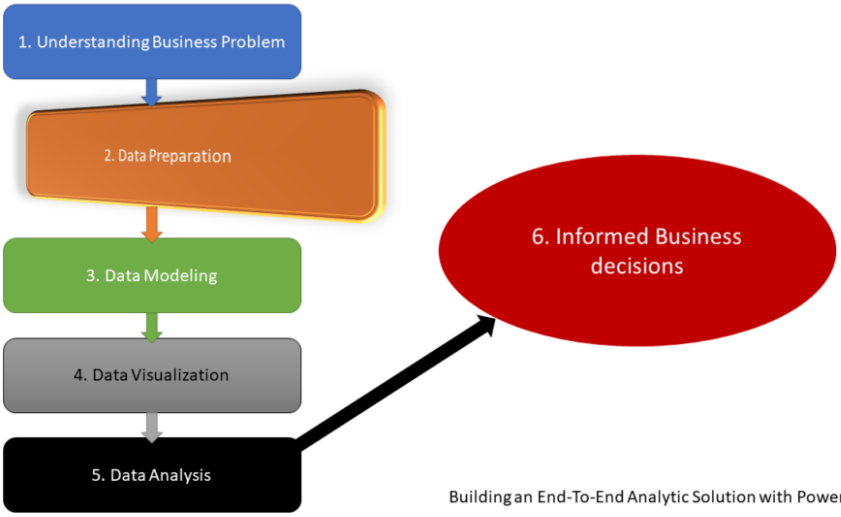
Pertanto, è estremamente importante identificare le domande chiave all'inizio, in modo da poter modellare e modellare i nostri dati per rispondere a tali domande nel modo più efficace.

Per il nostro set di dati, non abbiamo a che fare con le "classiche" domande di business, poiché non ci sono vendite, prodotti, promozioni... Tuttavia, non lo rende meno "degne", figuriamoci se possiamo permetterci di saltare alcuni dei passaggi sopra. Alcune delle nostre domande "commerciali" potrebbero essere:

- Quali sono i luoghi più rischiosi della città?
- Quale momento della giornata è il più critico?
- Qual è la percentuale di pedoni tra tutti i feriti?
- Quali sono i comuni della città con il più alto tasso di incidenti?
- Quali sono le tipologie di auto più frequentemente coinvolte negli incidenti?

L'obiettivo finale nel trovare le risposte a queste domande sarebbe identificare gli indicatori chiave che causano le collisioni (fase di analisi dei dati), e cercare in qualche modo di agire e prevenire incidenti futuri, o almeno ridurne il numero (prendere decisioni informate).

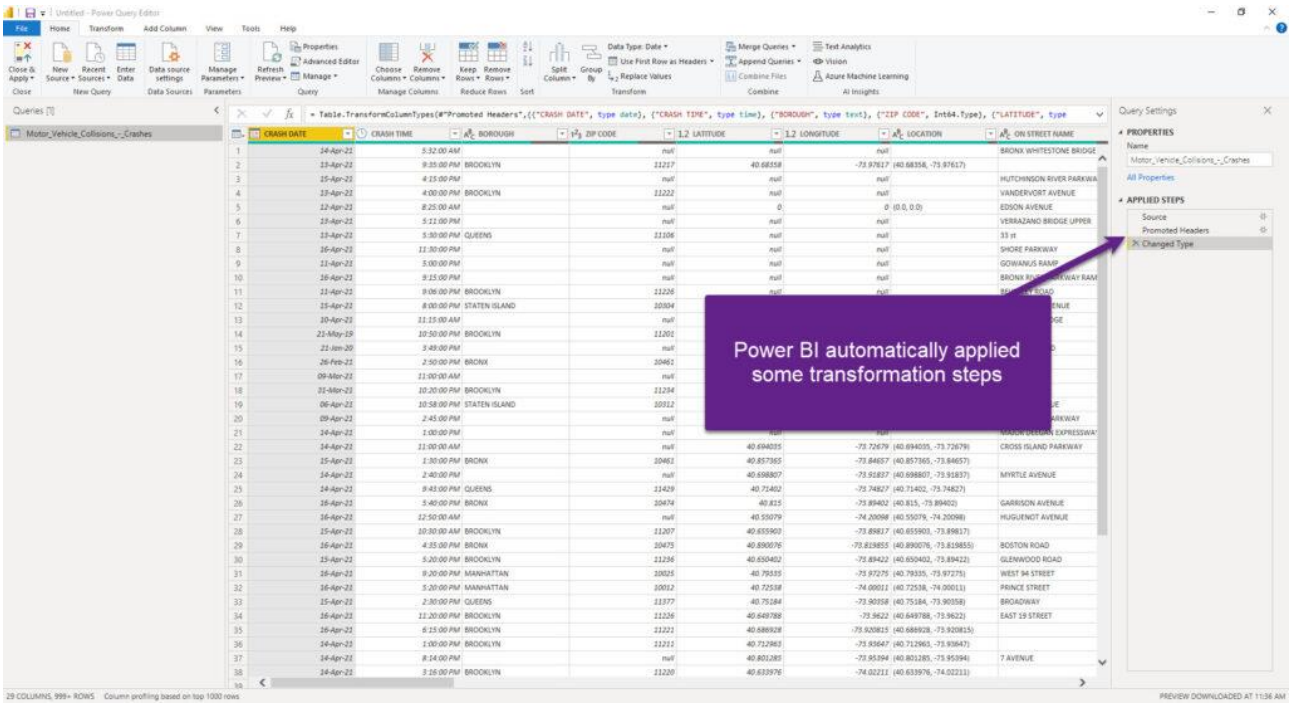
Power BI: Parte 2 – Data Preparation



Nella parte precedente, abbiamo posto alcune basi teoriche dietro il processo di costruzione di una soluzione analitica end-to-end e spiegato perché è di fondamentale importanza comprendere i problemi aziendali *PRIMA di* costruire una soluzione. Ora è il momento di tirarci su le maniche e iniziare a lavorare sul serio con il nostro set di dati. Come promemoria, utilizzeremo un set di dati aperto sulle collisioni di veicoli a motore a New York, che può essere trovato https://data-cityofnewyork-us.translate.goog/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95?_x_tr_sl=en&_x_tr_tl=it&_x_tr_hl=it&_x_tr_pto=nui

Prima guarda nel set di dati

I dati sono archiviati nel formato CSV e abbiamo una tabella piatta contenente circa 1,8 milioni di righe e 29 colonne. Diamo una rapida occhiata ai dati una volta importati in Power BI:

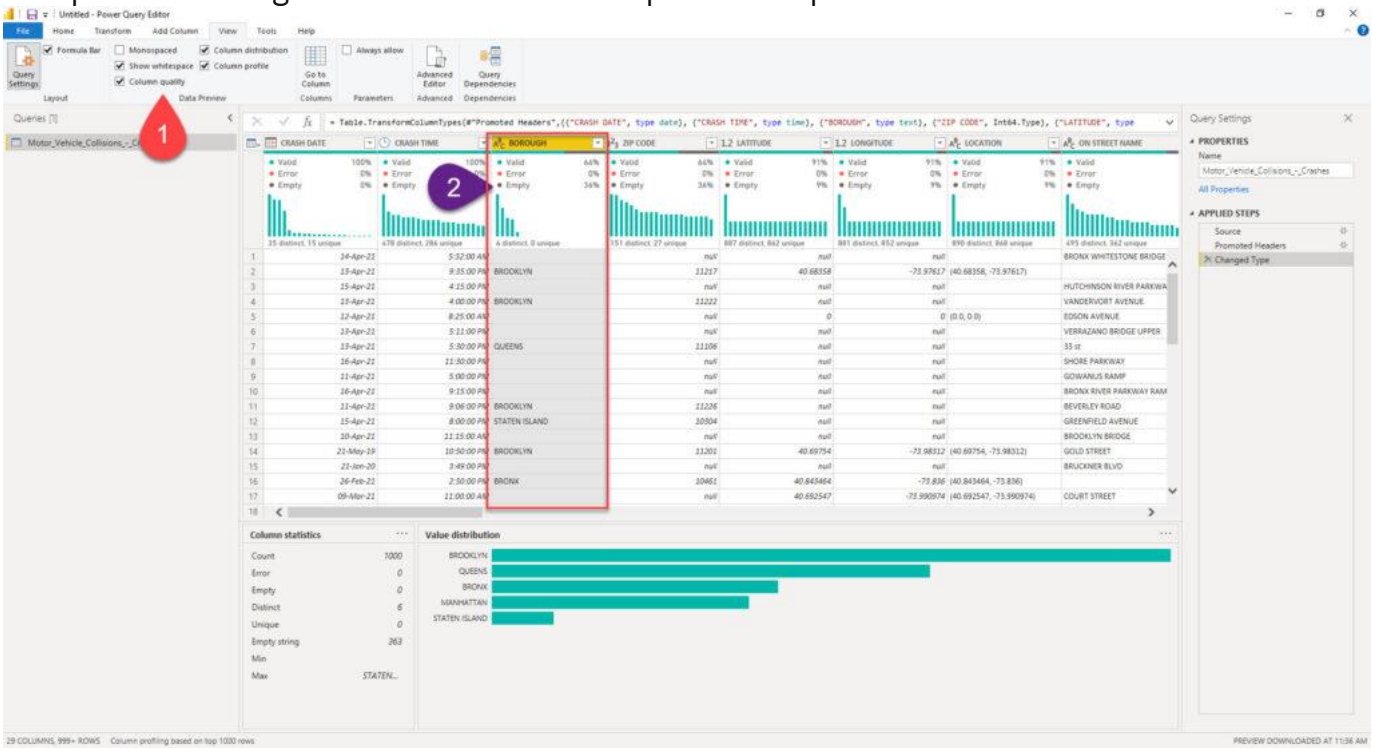


Preparazione dei dati

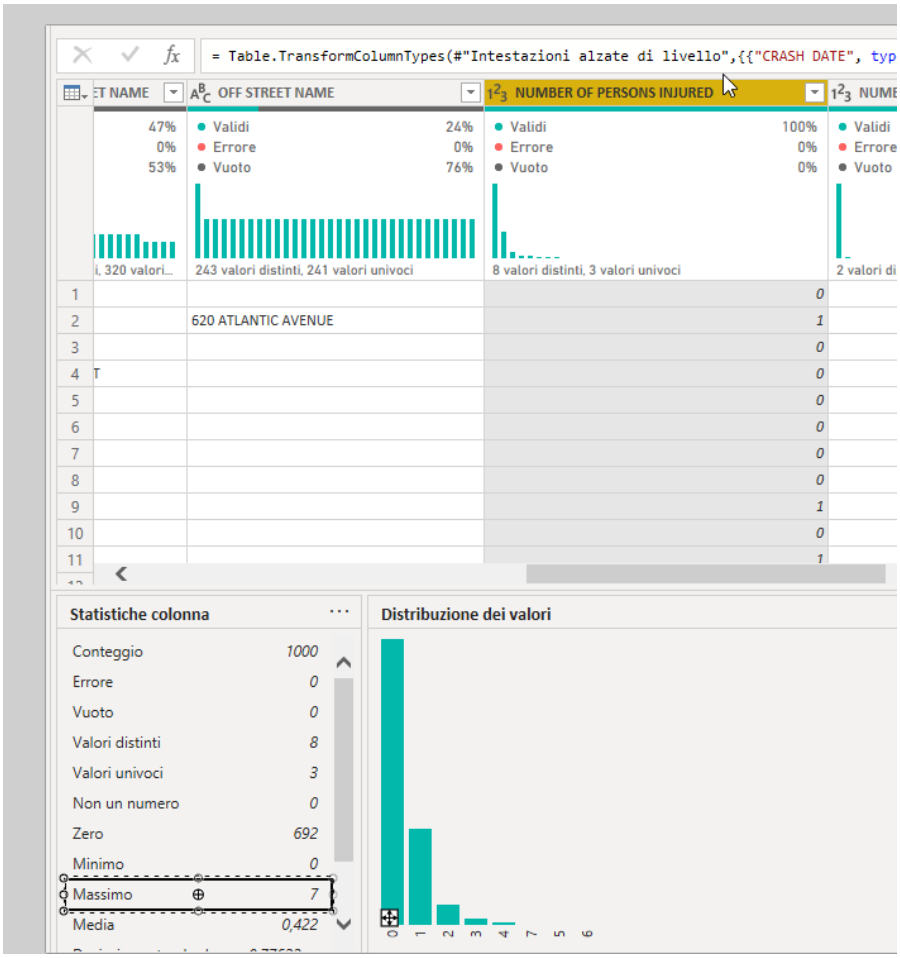
Poiché nel nostro esempio abbiamo a che fare con il file CSV, l'editor di Power Query è il posto più ovvio in cui applicare tutto il nostro lavoro di preparazione dei dati. Se dovessimo utilizzare, ad esempio, il database SQL come origine dati, avremmo potuto anche eseguire la modellazione dei dati sul lato sorgente, all'interno del database stesso!

Profilazione dei dati

Per cominciare, l'editor di Power Query offre un set di funzionalità molto utile per eseguire la profilatura dei dati . Andrò alla scheda Visualizza *e attiverò le* funzioni *Qualità colonna, Distribuzione colonna e Profilo colonna* per aiutarmi a comprendere meglio i dati e identificare potenziali problemi che devono essere risolti.



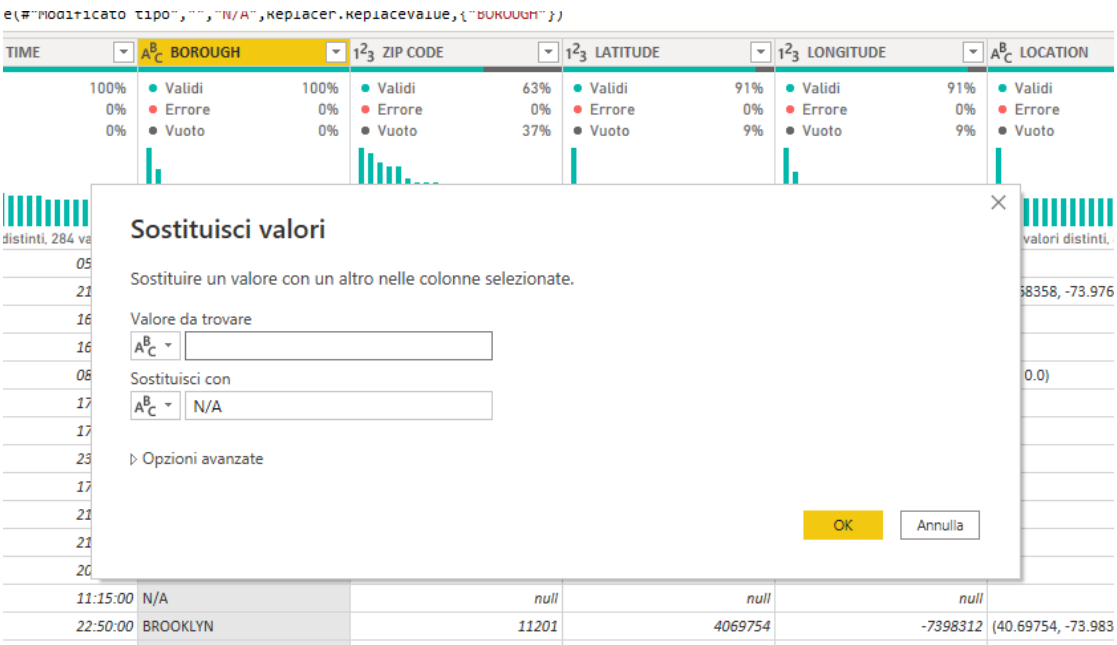
Ciò mi consentirà di individuare immediatamente che, ad esempio, è presente il 36% di valori mancanti per la colonna Borough. Sulla base dei risultati, posso decidere di lasciarlo così o applicare alcune trasformazioni aggiuntive per correggere i dati mancanti o incompleti. Ad esempio, posso decidere di sostituire tutti i valori vuoti o nulli con N/A o qualcosa di simile. Potrei anche identificare rapidamente valori anomali o anomalie (se presenti). Immaginiamo di profilare la colonna *Numero di Infortunati* :



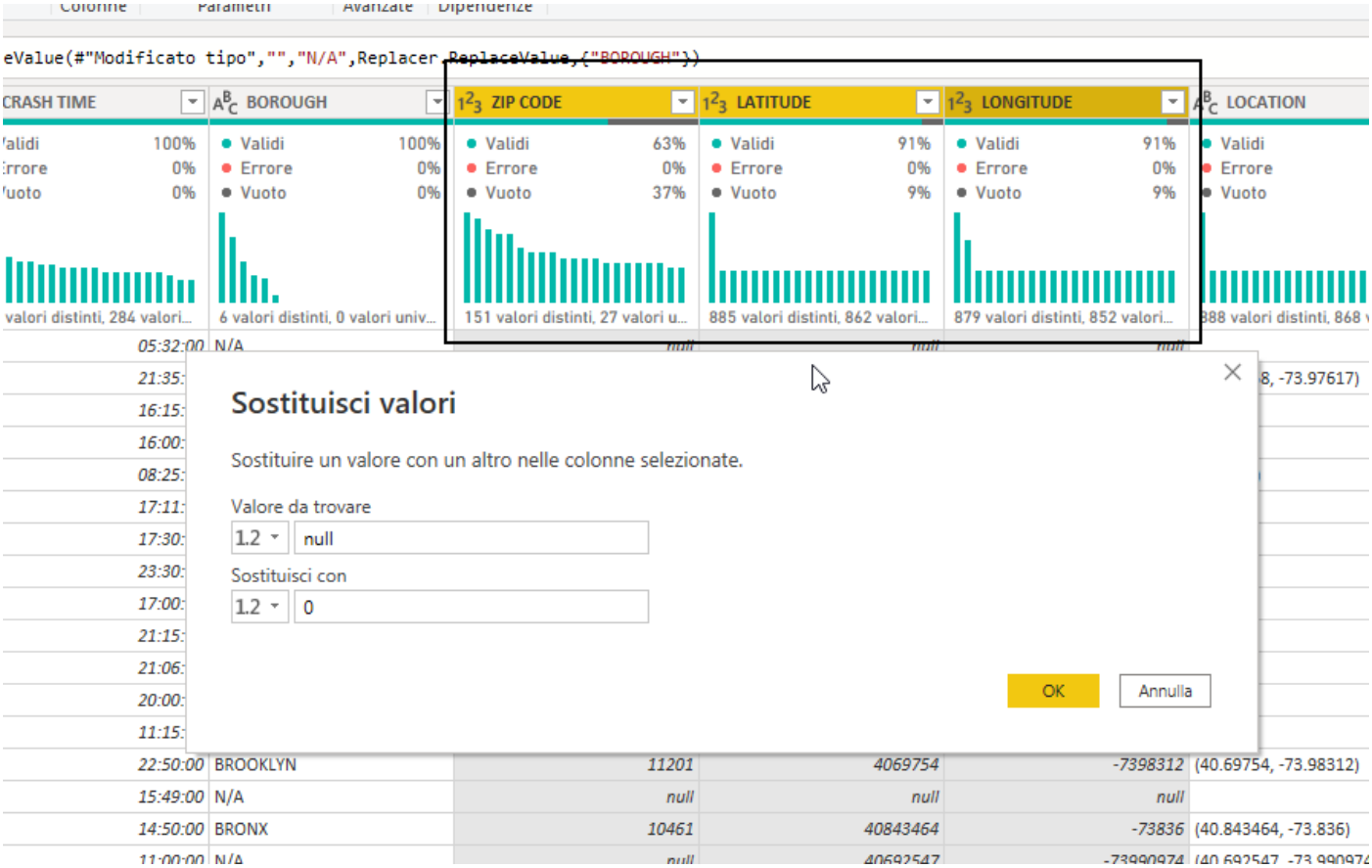
Se ci fossero delle anomalie nei dati (cioè invece di 7 per il numero massimo di feriti, diciamo 7000), saremmo in grado di individuarle subito e reagire di conseguenza!

Modellazione dei dati

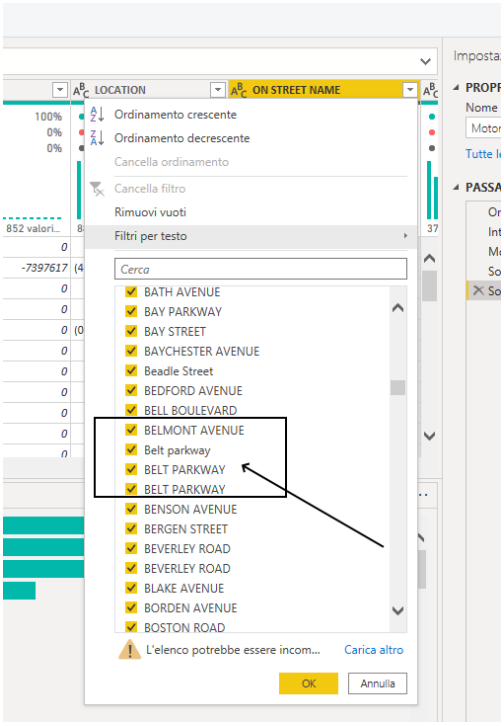
È tempo di migliorare il nostro set di dati e di investire un po' di impegno in più per migliorare la qualità dei dati. Iniziamo con la sostituzione dei valori vuoti con N/A nella colonna *Borough* h:



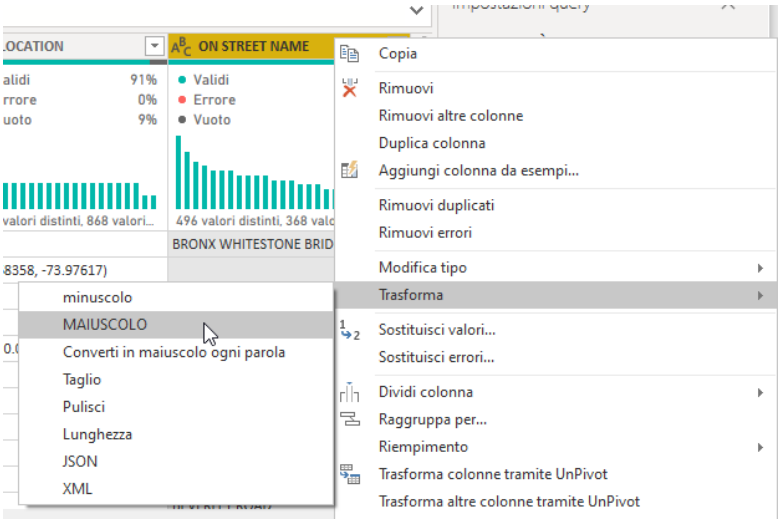
Il prossimo passo sarà pulire le colonne numeriche. CAP è la colonna del numero intero, mentre Latitudine e Longitudine sono rappresentate come valori decimali. Detto questo, sostituiremo i null con il valore 0 in ciascuna di queste colonne:



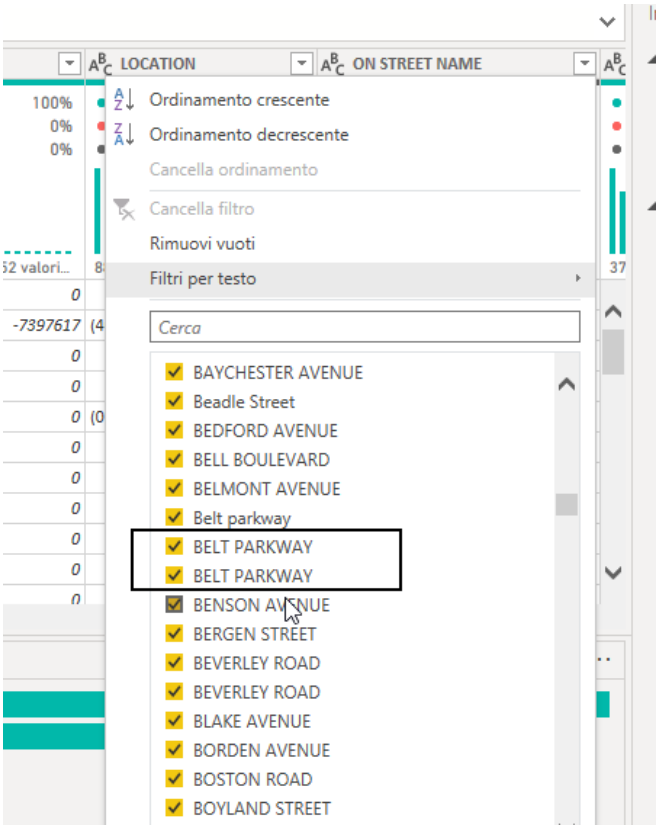
È stato facile e veloce, giusto? Ora andiamo avanti e proviamo a profilare altre colonne e controlliamo se sono necessarie alcune trasformazioni più sofisticate. La rubrica *Sul nome della via* è estremamente importante, perché serve per rispondere a una delle domande cruciali del business: quali sono i luoghi più rischiosi della città? Pertanto, dobbiamo assicurarci che questa colonna abbia il più alto livello di qualità dei dati.



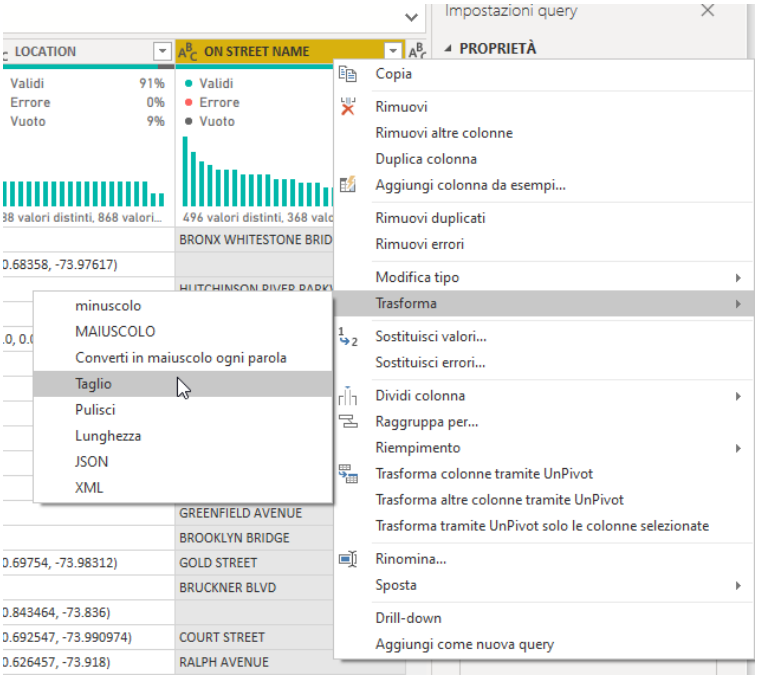
Aspetta!!!!! Belt Parkway è uguale a Belt Parkway, giusto? Ebbene, in realtà – Si Ma, nel linguaggio M di Power Query, la distinzione tra maiuscole e minuscole renderà queste due entità completamente diverse! Quindi, dobbiamo conformare i valori per essere in grado di ottenere risultati corretti nei nostri rapporti:



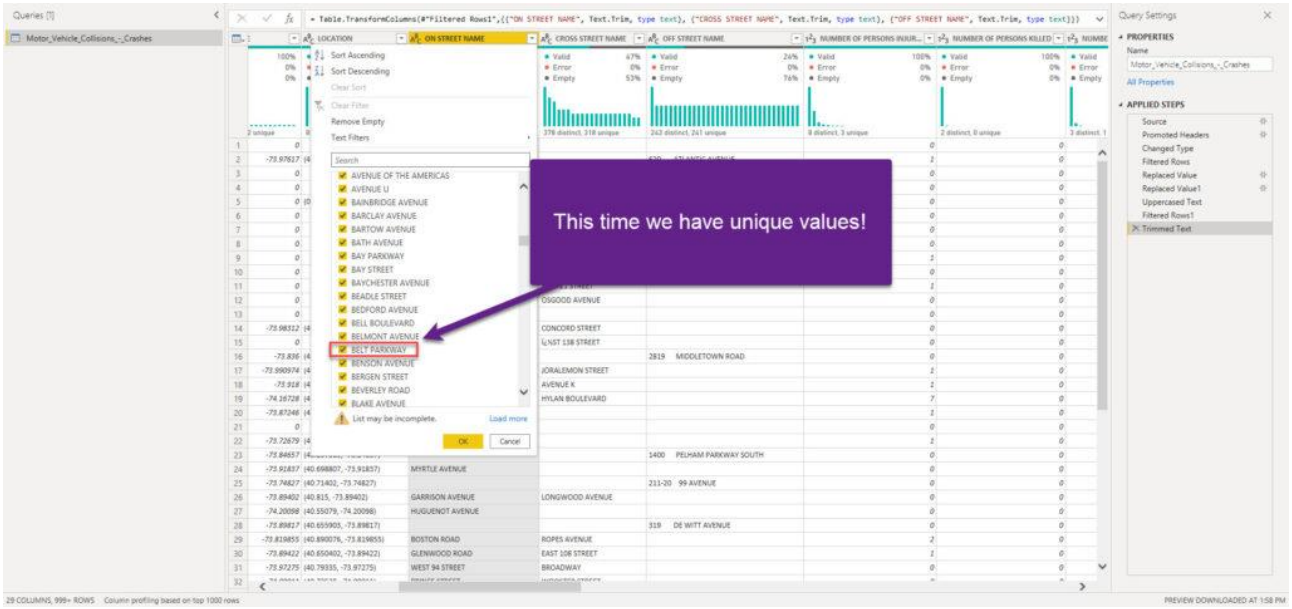
Come puoi vedere, applicherò la trasformazione maiuscola a tutte le colonne contenenti i nomi delle strade, e ora dovremmo essere a posto:



Perché abbiamo due esattamente gli stessi valori maiuscoli per BELT PARKWAY? Bene, il file CSV originale a volte può contenere caratteri nascosti, come tabulatore, nuova riga o spazio. Non preoccuparti, ho buone notizie per te: Power Query ti consente di risolvere questo problema specifico con un clic!



Questa volta abbiamo usato la trasformazione Trim per rimuovere o caratteri vuoti iniziali e finali. E, controlliamo di nuovo se questo ha risolto il nostro problema con i valori duplicati:

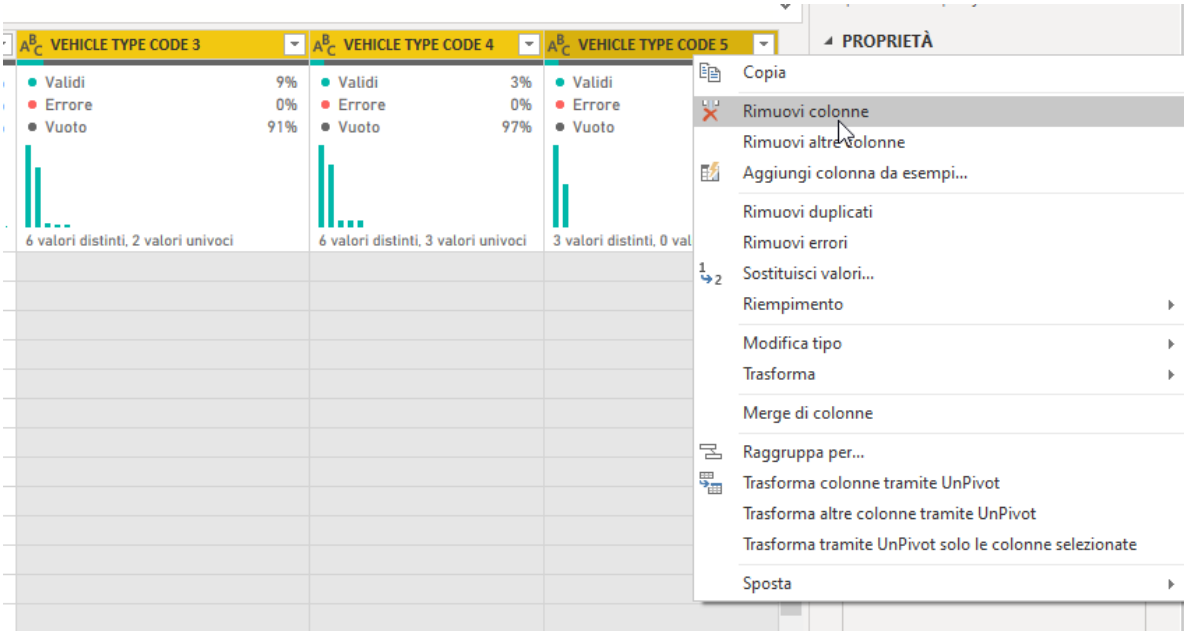


Infine, la nostra colonna appare come previsto: abbiamo valori univoci!

Pensando al futuro...

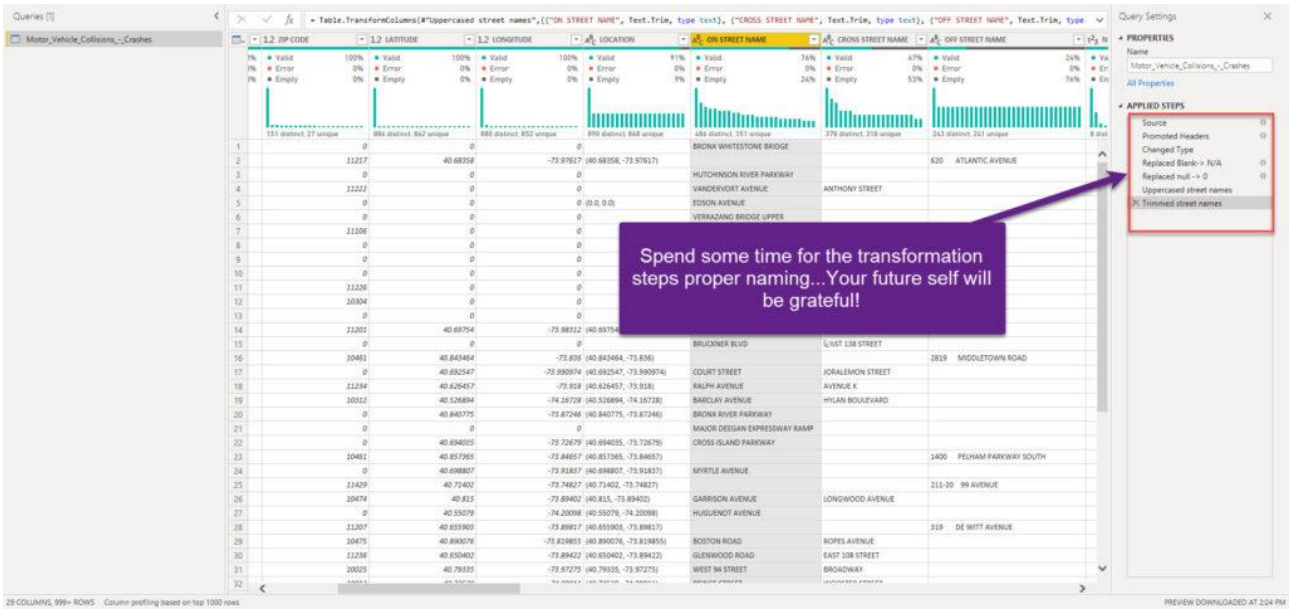
Ora puoi essere tentato di premere il pulsante *Chiudi e applica* e iniziare a creare belle visualizzazioni in Power BI. Tuttavia, sii paziente, poiché dobbiamo fare uno sforzo aggiuntivo prima di chiudere l'editor di Power Query.

Prima considerazione: abbiamo bisogno di tutte e 29 le colonne per la nostra soluzione analitica? Quindi, seguiamo le migliori pratiche relative all'ottimizzazione del modello di dati eliminiamo i dati non necessari. Ci sono 6 colonne con oltre il 90% di valori vuoti, quindi perché mai dovremmo gonfiare il nostro modello di dati con queste colonne quando non possono fornire qualche spunto utile?!



Ora sembra molto meglio! Prima di procedere alla fase successiva del nostro processo e iniziare a creare un modello di dati efficiente, è necessario eseguire un'altra operazione per rimanere in linea con le procedure consigliate quando si lavora con l'editor di Power Query .

Rinominerò ogni passaggio di trasformazione, in modo che se qualcuno (o anche io) apre questo file tra qualche mese, io sappia esattamente quale passaggio esegue quale trasformazione! Voglio dire, è facile quando hai solo pochi passaggi di trasformazione, ma una volta che ti trovi decine di passaggi di trasformazione, le cose diventano rapidamente più ingombranti ... invece di vedere ciascuno dei passaggi cercando di capire cosa fa, sarai in grado di cogliere facilmente la logica:



Fidati di me, il tuo futuro ti sarà estremamente grato dopo pochi mesi :)

Conclusione

Dopo aver sottolineato l'importanza di comprendere i problemi aziendali che devono essere risolti dalla soluzione analitica, in questa parte ci siamo sporcati le mani e

abbiamo iniziato a modellare i nostri dati per prepararli a rispondere a varie domande aziendali.

Durante il processo di preparazione dei dati, abbiamo eseguito la profilazione dei dati e identificato diversi problemi che potrebbero potenzialmente danneggiare la nostra soluzione finale, come valori mancanti o duplicati. Utilizzando lo strumento di trasformazione integrato di ***Power Query*** , siamo stati in grado di risolvere rapidamente le incoerenze dei dati e preparare il terreno per la fase successiva: la modellazione dei dati! Non dimenticare che Power Query Editor ti consente non solo di applicare trasformazioni complesse utilizzando una semplice interfaccia utente, senza alcuna competenza di codifica, ma ti offre anche la possibilità di migliorare significativamente il tuo modello di dati utilizzando linguaggio M se necessario.