W24D4 – Esercitazione finale

Ho caricato i file csv e ho iniziato la pulizia dei dati, correggendo il loro tipo ed eliminando le colonne non utili per l'analisi.

In products, ho sostituito con "Altro" le categorie di prodotti che in realtà erano null.

In order\_items ho aggiunto la colonna dei ricavi sommando price e freight.

In orders, ho considerato come data dell'ordine il campo order\_purchase\_timestamp (rinominata order\_date), questo perché ho visto che nel campo order\_approved\_at c'erano diversi valori null nonostante lo status dell'ordine fosse "delivered".

Inoltre, ho notato che c'erano 6 casi in cui risultava una data di consegna dell'ordine nonostante l'ordine stesso fosse segnalato come "canceled"; di conseguenza, ho gestito la situazione attraverso una nuova colonna "canceled\_flag", togliendo quei 6 casi sbagliati.

In seguito, ho fatto una doppia left-join per unire in un unico dataset orders, order\_items e order\_reviews.

Infine, nel dataset customers, ho dovuto aggiungere tramite join una colonna in cui è specificato il nome completo di ciascuno stato del Brasile, per far sì che funzioni al meglio la visualizzazione geografica.

Ho rinominato alcune colonne e poi sono passato alla fase di visualizzazione.