Data Transformation in Power Query

The dataset was extracted and imported into Power BI's Power Query for data validation and cleaning. The column profiling was changed from 'based on Top 1000 rows' to 'based on entire dataset'. The transformation done on each table is outlined below:

Olist_Order_items-: This has 112,650 rows and 7 columns. The shipping_limit timestamp which was a Date-time column was split by delimiter (space) into date and time columns and renamed accordingly. The datatype of the price and freight_value Columns were changed from decimal to fixed decimal.

Olist_Orders - This has 99,441 rows and 8 columns. The single Date-time columns were split by delimiter (space) into date and time columns and renamed accordingly. These includes order_purchase_timestamp, order_approved_at, order_delivered_carrier_date,

order_delivered_customer_date, and order_estimated_delivery_date.

The time Columns were then removed.

Olist_customer - This has 99,441 rows and 5 columns. The customer_city column was changed from lower case to proper case by capitalising each word. Special characters were removed from the customer_city column. The abbreviated state names on the customer_state column was replaced with their full names. Columns were then trimmed to remove any leading or trailing whitespaces.

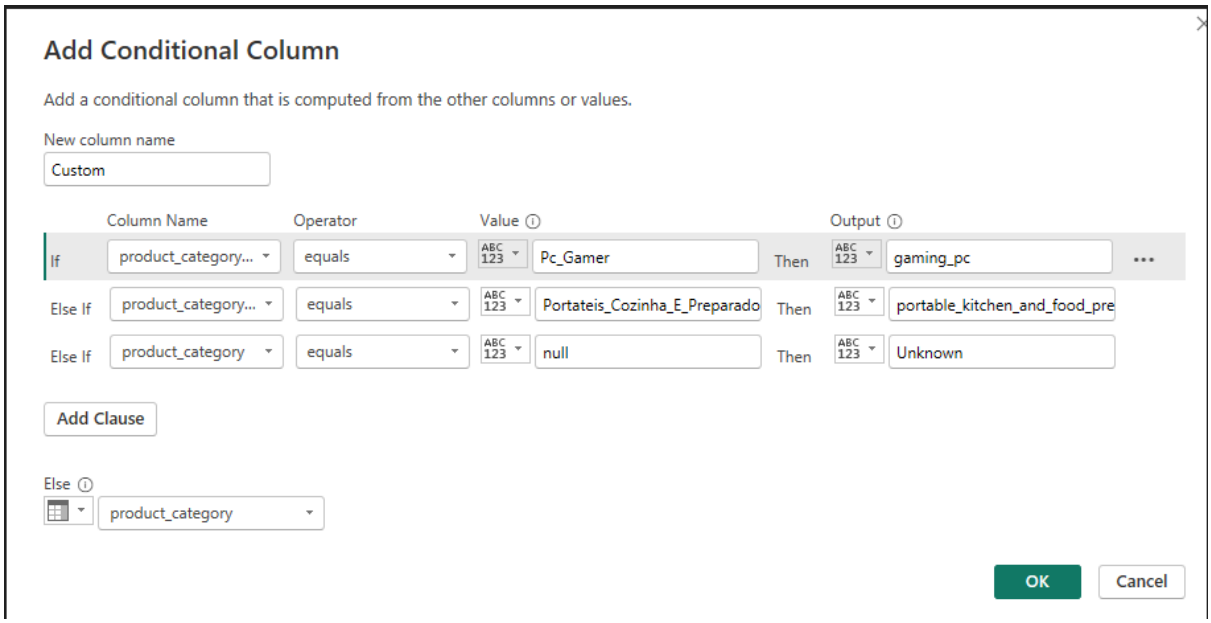
Olist_geolocation - This has 1,000,163 rows and 5 Columns. The first row was made to be the header as the headers were showing column 1, column 2, etc., after it was imported. Duplicates were then removed from the geolocation_zip_code_prefix column. Special characters were removed from the geolocation_city column. The abbreviated state names on the geolocation_state column was replaced with their full names. Columns were then trimmed to remove any leading or trailing whitespaces.

Olist_order_payments - This contains 103,886 rows and 5 columns. The payment_value column was changed from decimal to fixed decimal.

Olist_order_reviews - This contains 100,000 rows and 7 columns. The review_creation_timestamp and review_answer_timestamp columns which were Date-time columns was split by delimiter (space) into date and time columns and renamed accordingly. The time Columns were then removed.

Olist_products - This had 32,951 rows, 9 Columns. The product_category_name column was changed from lower case to proper case by capitalising each word. The product category_name_translation table (having 72 rows & 2 columns) was merged with the product table, the resulting added table-column was then expanded to select needed column. The new column was renamed accordingly.

However, I realized that 3 categories didn't have their English translation available and was blank. I found the translation by research, created a custom column with an if-then/else, statement that allowed for recreation of the column with the blank rows now filled accordingly.



Similarly, one category had the same name in column for Brazil name variant and English name variant. This was replaced with the English translation (From "La_Cuisine" to "Kitchen_Supplies". Columns like product_height_cm, product_width_cm, and the likes, which were not needed for the current analysis were removed.

Olist_sellers - This has 3,095 rows and 4 columns. Special characters were removed from the seller_city column. The abbreviated state names on the seller_state column was replaced with their full names. Columns were then trimmed to remove any leading or trailing whitespaces.