# Data Table Schema

## retailer_pricing

Dynamic product pricing data from 6 competing US retailers, from August 2019 to March 2020. *~22.5 million rows & 11 columns.* Size: ~1GB zipped, ~3GB unzipped. Source: not public.

| Field | Type | Description |
|---|---|---|
| **store** | STRING | The retailer under which this product is listed. (0=amazon.com, 1=zappos.com, 2=macys.com, 3=neimanmarcus.com, 4=saksfifthavenue.com, 5=bloomingdales.com) |
| **title** | STRING | The title of the product listing. |
| **sku** | STRING | The retailer-specific stock-keeping unit (SKU). |
| **groupid** | STRING | The group-level SKU, identifying the same or similar product listing across different retailers. |
| **brand** | STRING | The brand of the product listing. |
| **color** | STRING | The color of the product listing. |
| **sizing** | STRING | The sizing of the product listing. |
| **category** | STRING | The category of the product listing. |
| **subcategory** | STRING | The subcategory of the product listing. |
| **price** | FLOAT | The price of the product, in USD. |
| **date** | STRING | The timestamp for when the listing details were recorded (format: yyyy-mm-dd). |

## amazon_reviews (2)

Reviews and product metadata of millions of Amazon products, from 1996 - 2018. We only provided data from the “All Beauty” category due to its size and high degree of overlap with the categories of the *retailer_pricing* dataset. Note that product IDs do not match between datasets.

Size: ~45MB zipped, ~120MB unzipped. Source.

If you would like to extend your analysis to other categories (e.g. “Clothing Shoes and Jewelry” is highly relevant), you can freely retrieve this data here after filling out a form. Note that this data will require some cleaning - here is a colab notebook that helps you get started.

**beauty_reviews**. *~370,000 rows & 9 columns.*

| Field | Type | Description |
|---|---|---|
| **overall** | FLOAT | Rating of the product |
| **verified** | BOOL | Whether it is verified that the user bought the product. |
| **reviewTime** | STRING | The timestamp for the review (format: yyyy-mm-dd). |

| | | |
|---|---|---|
| **reviewerID** | STRING | ID of the reviewer |
| **asin** | STRING | ID of the product |
| **reviewText** | STRING | Text of the review |
| **summary** | STRING | The summary of the review |
| **vote** | FLOAT | Number of people who found the review helpful |
| **style** | DICT | A dictionary of the product metadata |

**beauty_metadata**. *~33,000 rows & 9 columns.*

| Field | Type | Description |
|---|---|---|
| **title** | STRING | The name of the product |
| **brand** | STRING | The name of the brand. |
| **rank** | FLOAT | The rank of the product in the Beauty category. |
| **asin** | STRING | The ID of the product |
| **description** | LIST | The description of the product. |
| **also_view** | LIST | Related products that other customers viewed. |
| **also_buy** | LIST | Related products that other customers bought. |
| **price** | FLOAT | The price of the product in USD |
| **similar_item** | LIST | Similar product table. |

## online_consumer_behavior (2)

Online event-based consumer behavior data from RetailRocket, collected over 4.5 months. Please refer to the source for extended details about the dataset.

Size: ~300MB zipped, ~1GB unzipped. Source

**events**. *~2.7 million rows & 5 columns.*

| Field | Type | Description |
|---|---|---|
| **timestamp** | INT | Unix timestamp for when the event occurred. |
| **visitorid** | INT | Unique identifier of the visitor |
| **event** | STRING | Event type ('view', 'addtocart', 'transaction') |
| **itemid** | INT | Unique identifier of the item |
| **transactionid** | FLOAT | Unique identifier of the transaction |

**item_properties**. *~20 million rows, 4 columns*.

| Field | Type | Description |
|---|---|---|
| **timestamp** | INT | Unix timestamp for when the event occurred. |
| **itemid** | INT | Unique identifier of the item. |
| **property** | STRING | Property of the item being recorded. All values were hashed except for *available* (1 = yes, otherwise 0) and *categoryid*, which indicates the item category identifier (refer to *category_tree.csv* for child/parent relationships between categoryid's) |
| **value** | STRING | Property value of the item. Refer to the source for extended details about how this is coded. |

## UK_retail_transactions

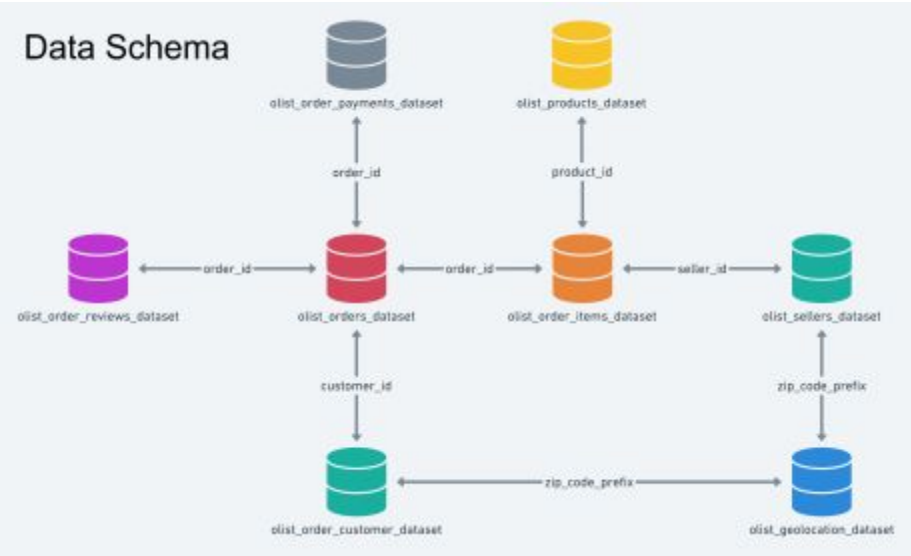
Transactions dataset from a UK-based online retailer, from 2009 - 2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.
*~1 million rows & 8 columns.* Size: ~15MB zipped, ~100MB unzipped. Source

| Field | Type | Description |
|---|---|---|
| **invoice_no** | STRING | A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation. |
| **stock_code** | STRING | A 5-digit integral number uniquely assigned to each distinct product. |
| **description** | STRING | Product name. |
| **quantity** | INT | The quantities of each product per transaction. |
| **invoice_date** | STRING | The day and time when a transaction was generated (format: mm-dd-yy HH:MM). |
| **unit_price** | FLOAT | Product price per unit in sterling (pounds). |
| **customerid** | FLOAT | A number uniquely assigned to each customer. |
| **country** | STRING | The name of the country where a customer resides. |

## BR_retail_transactions (8)

Transactions dataset containing 100k orders made at multiple marketplaces in Brazil through the Olist Store, from 2016 - 2018. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and customer reviews. Refer to the source for extended details on the data schema.

Size: 45MB zipped, 120MB unzipped. Source



**olist_customers_dataset**. *~100,000 rows & 5 columns.*

| Field | Type | Description |
|---|---|---|
| **customer_id** | STRING | Keys to the orders dataset. Each order has a unique customer_id. |
| **customer_unique_id** | STRING | Unique identifier of a customer. |
| **customer_zip_code_prefix** | INT | First five digits of customer zip code |
| **customer_city** | STRING | Customer city name. |
| **customer_state** | STRING | Customer state. |

**olist_geolocation_dataset**. *~1 million rows & 5 columns.*

| Field | Type | Description |
|---|---|---|
| **geolocation_zip_code_prefix** | INT | First five digits of zip code. |
| **geolocatation_lat** | FLOAT | Associated latitude. |
| **geolocation_lng** | FLOAT | Associated longitude. |
| **geolocation_city** | STRING | Name of the associated city. |
| **geolocation_state** | STRING | Name of the associated state. |

**olist_order_items_dataset**. *~100,000 rows & 7 columns.*

| Field | Type | Description |
|---|---|---|
| **order_id** | STRING | Order unique identifier. |
| **order_item_id** | INT | Sequential number identifying number of items included in the same order. |
| **product_id** | STRING | Product unique identifier. |
| **seller_id** | STRING | Seller unique identifier. |
| **shipping_limit_date** | STRING | Shows the seller shipping limit date for handling the order over to the logistic partner. |
| **price** | FLOAT | Item price (BRL). |
| **freight_value** | FLOAT | Item freight value item (if an order has more than one item the freight value is split between items) |

**olist_order_payments_dataset**. *~100,000 rows & 5 columns.*

| Field | Type | Description |
|---|---|---|
| **order_id** | STRING | Unique identifier of an order. |
| **payment_sequential** | INT | Customers may pay with more than one payment method. If they do so, a sequence is created to accommodate all payments |
| **payment_type** | STRING | Method of payment chosen by the customer. |
| **payment_installments** | INT | Number of installments chosen by the customer. |
| **payment_value** | FLOAT | Transaction value (BRL). |

**olist_order_reviews_dataset**. *~100,000 rows & 7 columns.*

| Field | Type | Description |
|---|---|---|
| **review_id** | STRING | Unique review identifier |
| **order_id** | STRING | Unique order identifier |
| **review_score** | INT | Note ranging from 1 to 5 given by the customer on a satisfaction survey. |

| | | |
|---|---|---|
| **review_comment_title** | STRING | Comment title from the review left by the customer, in Portuguese. |
| **review_comment_message** | STRING | Comment message from the review left by the customer, in Portuguese. |
| **review_creation_date** | STRING | Shows the date in which the satisfaction survey was sent to the customer. |
| **review_answer_timestamp** | STRING | Shows satisfaction survey answer timestamp. |

**olist_orders_dataset**. *~100,999 rows & 8 columns.*

| Field | Type | Description |
|---|---|---|
| **order_id** | STRING | Unique identifier of the order. |
| **customer_id** | STRING | Key to the customer dataset. Each order has a unique customer_id. |
| **order_status** | STRING | Reference to the order status (delivered, shipped, etc). |
| **order_purchase_timestamp** | STRING | Shows the purchase timestamp. |
| **order_approved_at** | STRING | Shows the payment approval timestamp. |
| **order_delivered_carrier_date** | STRING | Shows the order posting timestamp. When it was handled to the logistic partner. |
| **order_delivered_customer_date** | STRING | Shows the actual order delivery date to the customer. |
| **order_estimated_delivery_date** | STRING | Shows the estimated delivery date that was informed to the customer at the purchase moment. |

**olist_products_dataset**. *~32,000 rows & 9 columns.*

| Field | Type | Description |
|---|---|---|
| **product_id** | STRING | Unique product identifier |
| **product_category_name** | STRING | Root category of product, in Portuguese. |
| **product_name_length** | FLOAT | Number of characters extracted from the product name. |
| **product_description_length** | FLOAT | Number of characters extracted from the product description. |
| **product_photos_qty** | FLOAT | Number of product published photos |
| **product_weight_g** | FLOAT | Product weight measured in grams. |
| **product_length_cm** | FLOAT | Product length measured in centimeters. |
| **product_height_cm** | FLOAT | Product height measured in centimeters. |
| **product_width_cm** | FLOAT | Product width measured in centimeters. |

**olist_sellers_dataset**. *~3,000 rows & 4 columns.*

| Field | Type | Description |
|---|---|---|
| **seller_id** | STRING | Seller unique identifier |
| **seller_zip_code_prefix** | INT | First 5 digits of seller zip code |
| **seller_city** | STRING | Seller city name |
| **seller_state** | STRING | Seller state |